Représentations gros-grain pour la modélisation des protéines : Propriétés mécaniques et interactions

Sacquin-Mora Sophie

Mémoire présenté en vue de l'obtention de l'

HABILITATION à DIRIGER des RECHERCHES

UFR Sciences du Vivant, Université Paris 7-Denis Diderot

par

Sophie Sacquin-Mora

Représentations gros-grain pour la modélisation des protéines :

Propriétés mécaniques et interactions.

Soutenue le 13 décembre 2011 devant le jury composé de :

Rapporteurs :	Anne-Claude Camproux
	Annick Dejaegere
	Yves-Henri Sanejouand
Examinateurs :	Anne Imberty
	Richard Lavery
	Martin Zacharias

Représentations gros-grain pour la modélisation des protéines : Propriétés mécaniques et interactions.

Sophie Sacquin-Mora

Chargée de recherche au CNRS

Laboratoire de Biochimie Théorique, CNRS UPR9080

Institut de Biologie Physico-Chimique

13 rue Pierre et Marie Curie, 75005 Paris

Résumé :

Je travaille au Laboratoire de Biochimie Théorique de l'Institut de Biologie Physico-Chimique, où j'ai été recrutée en octobre 2006 comme chargée de recherche CNRS suite à trois ans de post-doctorat et une thèse de physico-chimie. Mes travaux de recherche portent sur l'utilisation de modèles gros-grain et le développement d'algorithmes pour l'étude des propriétés mécaniques des protéines et des interactions protéine-protéine.

Sur le plan mécanique, j'ai développé le programme ProPHet (Probing Protein Heterogeneity), qui permet de sonder la rigidité protéique à l'échelle du résidu et d'étudier la réponse d'un système moléculaire soumis à une déformation anisotrope. Cette réponse mécanique peut être mise en rapport avec les propriétés structurales de la protéine concernée (notamment l'agencement de ses différents éléments de structure secondaire), mais aussi avec son fonctionnement biologique. Au cours des dernières années, ce logiciel a été exploité de manière autonome ou bien dans le cadre de collaborations avec des groupes expérimentaux français ou internationaux. ProPHet peut également être utilisé en complément d'approches de modélisation tout-atome classiques (comme la dynamique moléculaire ou la métadynamique), afin d'apporter des informations complémentaires sur la mécanique d'un système protéique.

En ce qui concerne les interactions protéiques, j'ai mis au point le programme MAXDo (Molecular Association via Cross-Docking) qui permet de mieux appréhender la spécificité des phénomènes de reconnaissance protéique via des calculs de docking-croisé à grande échelle. Ce programme, qui a été développé dans le cadre du projet Decrypthon (www.decrypthon.fr), a tout d'abord été exploité sur une grille de calcul universitaire, avant de faire l'objet d'un dossier de valorisation et d'être transposé sur la grille d'internautes World Community Grid (www.worldcommunitygrid.org) mise en place par IBM, sous le nom de projet Help Cure Muscular Dystrophy (HCMD). Ce projet en est actuellement à sa seconde phase, où MAXDo est exploité en association avec le programme de prédiction des interfaces protéiques JET (Joint Evolutionary Trees), qui a été développé en collaboration avec l'équipe de bioinformatique du Pr. Alessandra Carbone (CNRS-Université Paris 6).

Mots-clés : Modélisation moléculaire, mécanique des protéines, interactions protéiques, réseau élastique, docking.

Table des matières

Première partie

Dossier Scientifique

Chapitre 1

Introduction générale

Les protéines sont des acteurs majeurs de l'ensemble des processus biologiques qui les font intervenir à des échelles spatiales variées, partant de quelques ångströms (la taille d'un site catalytique) pour arriver au niveau cellulaire, et couvrant des échelles de temps allant de la femtoseconde à l'heure[1]. Dans ce paysage spatio-temporel, les simulations à l'échelle tout-atome ont fait leur preuve pour décrire avec précision des phénomènes impliquant des systèmes protéiques de plusieurs centaines d'acides aminés sur des durées allant jusqu'à la microseconde,

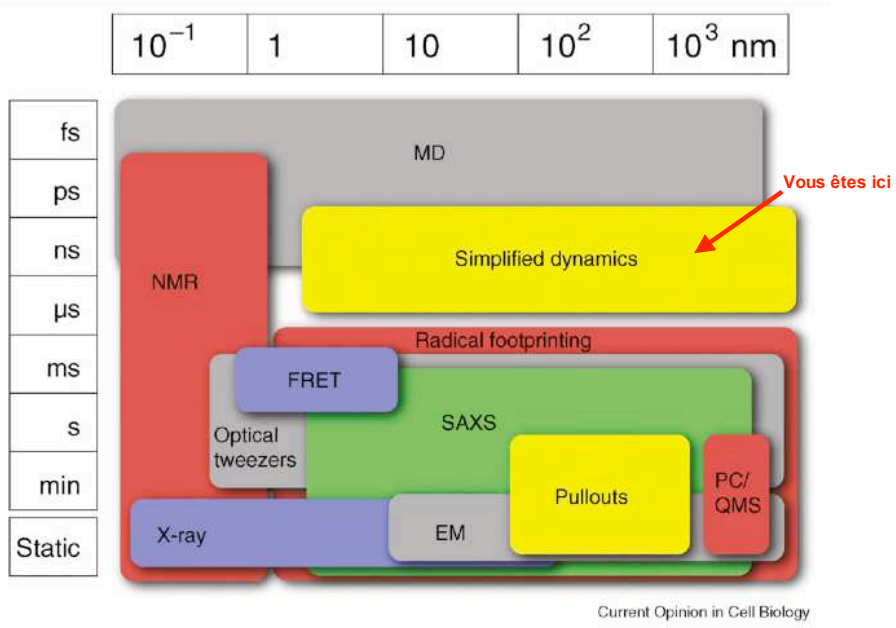


FIG. 1.1 – *Un panorama spatial et temporel de différentes méthodes, expérimentales ou théoriques, disponibles pour étudier les systèmes biologiques macromoléculaires. Les modèles gros-grain sont à ranger dans la catégorie « Simplified dynamics ». Figure issue de [1] .*

voir la figure 1.1. Au delà de ces frontières, il s'est avéré nécessaire d'avoir recours à ce que l'on nomme couramment les modèles « gros-grain », où l'élément de base pour la représentations du système protéique n'est plus l'atome seul mais un « pseudo-atome » de taille variable, voir la figure 1.2, ce qui va permettre de réduire de manière significative le nombre de degrés de liberté et donc la complexité du système étudié. Ces modèles gros-grain ont actuellement plus de trente-cinq années de développement derrière eux et nous n'évoquerons donc dans cette introduction que les jalons qui ont mené au travail présenté ici, le lecteur curieux pouvant toujours se référer à des revues récentes publiées sur le sujet [2, 3, 4]. En permettant d'accéder à des tailles de systèmes et des durées inenvisageables dans le cadre de simulations tout atome, ces représentations ont su faire la preuve de leur utilité et constituent désormais un élément indispensable de la « trousse à outils » du modélisateur moléculaire, qui peut les utiliser seules, ou en association avec d'autres modèles dans le cadre de simulations multi-échelle [5, 6, 7, 8, 9, 10, 11]

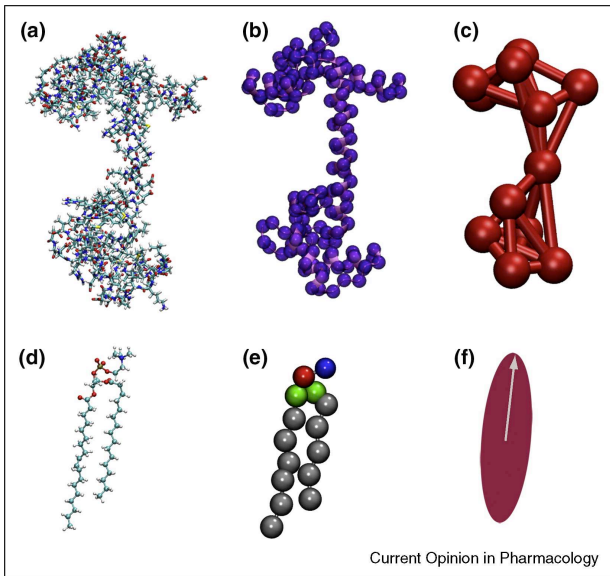


FIG. 1.2 – *Différents niveaux de représentations gros-grain. Calmoduline,(a) en tout atome, (b) avec un pseudo-atome par résidu, (c) avec douze noeuds pour la protéine. Lipide POPC (d) en tout atome, (e) avec le modèle MARTINI [12], (f) en ellipsoïde de Gay-Bern. Figure issue de [4] .*

1.1 Une (très) brève histoire des représentations gros-grain pour les protéines

1.1.1 Les potentiels empiriques pour le repliement protéique

Les premières réprésentations réduites ont été développées pour modéliser le repliement protéique[13]. Un des exemples les plus connus, le modèle de *Gō*[14], décrit la protéine comme une chaîne d'acides aminés, représentés chacun par un pseudo-atome, et dont la structure est biaisée en faveur de sa configuration native. Bien que très simple, cette représentation permet de reproduire divers aspects thermodynamiques et cinétiques du processus de repliement protéique[15, 16], la surface d'énergie de la protéine adoptant la forme classique en « enton-

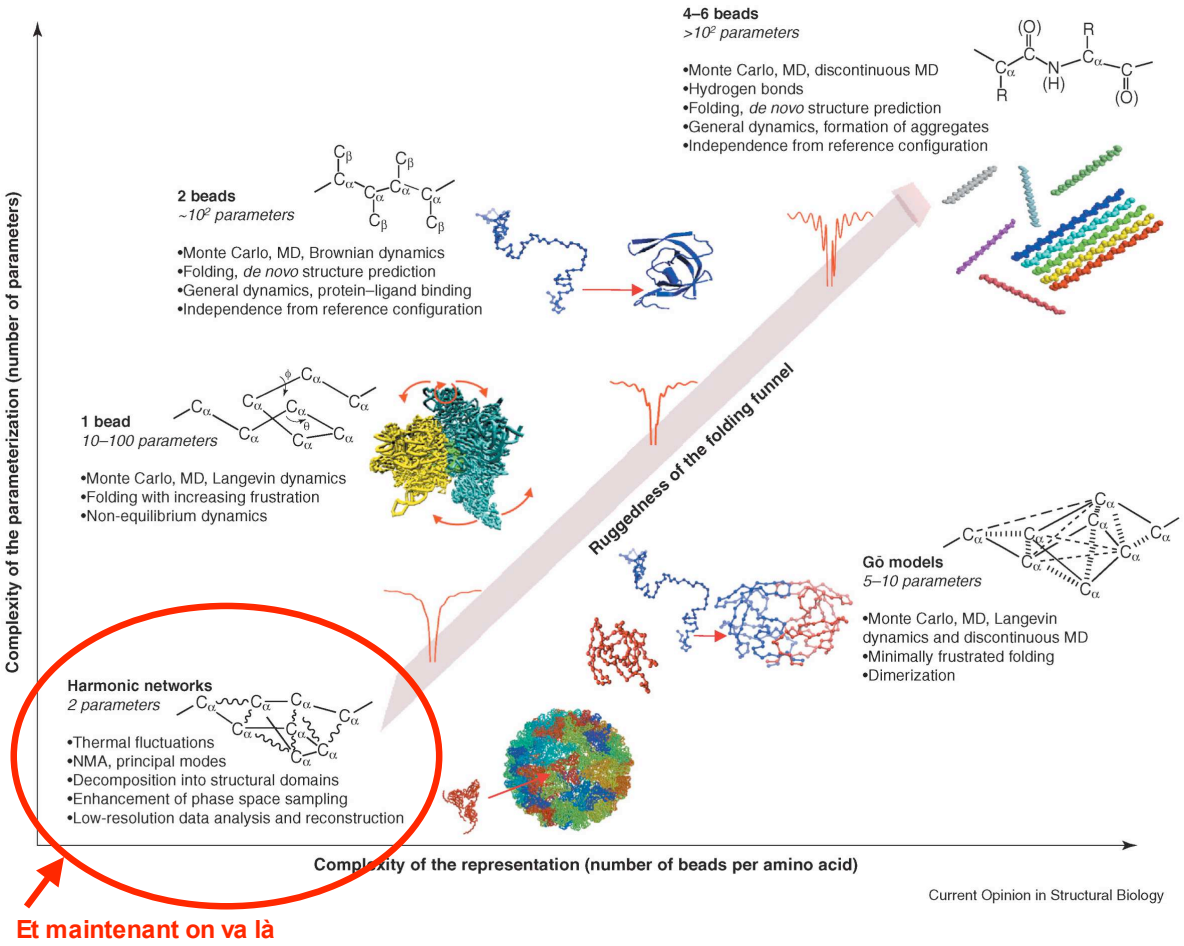


FIG. 1.3 – *Un échantillon des représentations gros-grain développées pour la modélisation des protéines. Figure issue de [2] .*

noir », qui permet de converger vers la structure native, voir la figure 1.3. Les modèles de *Gō* présentent des potentiels de type « knowledge based », dont la paramétrisation exploite les données structurales disponibles dans la Protein DataBank[17]. Les potentiels de Miyazawa et Jernigan[18] appartiennent également à cette catégorie. L'énergie de contact entre résidus est alors paramétrisée à partir de la distribution de ces contacts dans l'ensembles des structures natives qui ont été déterminées expérimentalement.

1.1.2 Le modèle en réseau de ressorts ou Gaussian Network Model

L'élaboration d'un modèle gros-grain passe par la définition des pseudo-atomes d'une part, et par la construction d'un potentiel d'interaction entre les particules d'autres part[4]. Les simulations tout atome utilisent en général un champ de force qui se partage entre termes covalents, comprenant les énergies de déformation pour les liaisons, angles et dièdres, et termes non-covalents, qui rendent compte des interactions électrostatiques et de van der Waals. Ce type de champ de force fait intervenir un grand nombre de constantes qu'il va falloir paramètriser de manière à obtenir un potentiel d'interaction qui soit transférable d'un système à l'autre.

En 1996, M. Tirion propose de donner une forme quadratique à l'énergie intramoléculaire

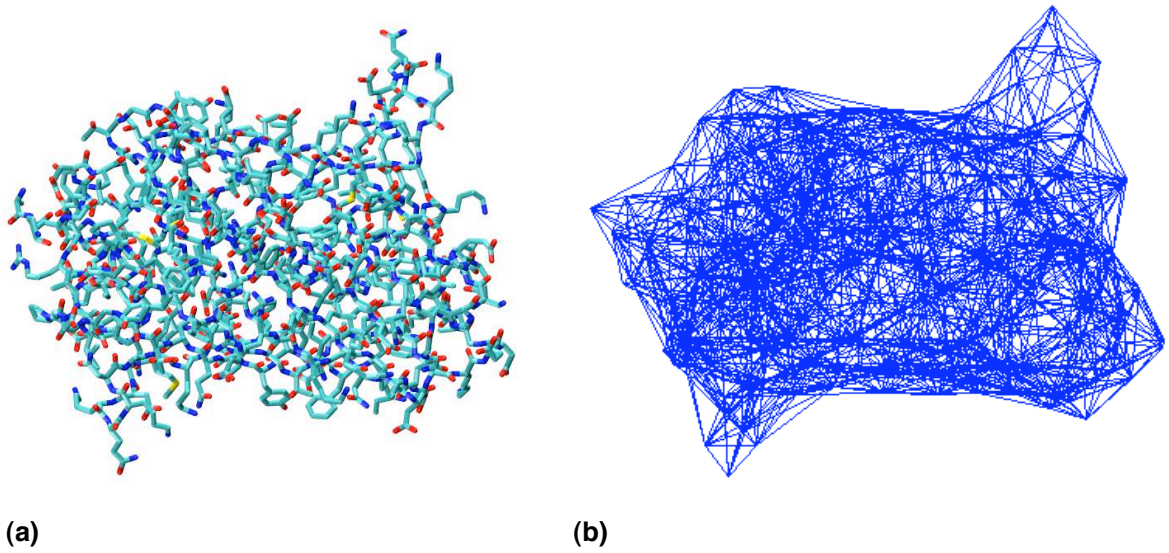


FIG. 1.4 – *Green Fluorescent Protein : (a) Représentation tout atome, (b) Réseau élastique construit à partir des carbones α situés à moins de $10\mathring{A}$.*

d'un système protéique[19], qui s'exprime alors très simplement selon l'équation suivante :

$$E = \frac{1}{2} \sum_{r_{ij}^0 < R_{cut}} \gamma \left(r_{ij} - r_{ij}^0 \right)^2 \tag{1.1}$$

Ce potentiel, qui est déterminé par les distances r_{ij} entre atomes dans la structure de référence du système (le plus souvent une structure cristallographique) ne dépend alors plus que de deux paramètres : Le rayon de coupure R_{cut} (typiquement compris entre 8[20] et 16 Å[21]) en deça duquel on considère que deux atomes i et j interagissent via un ressort harmonique, et γ, la constante de force de ce ressort, qui est la même pour toutes les paires de pseudo-atomes en interaction, voir la figure 1.4. Dans son article fondateur[19], M. Tirion montre comment ce modèle simplifié à l'extrême permet néanmoins de reproduire efficacement les modes de vibration basse-fréquence des protéines et les fluctuations atomiques qu'expriment les facteurs de température (ou facteurs de Debye-Waller) obtenus par cristallographie. Cette étude initiale va connaître ensuite une descendance foisonnante (en mars 2011, l'article original de 1996 était cité plus de cinq cents fois sur la base de données ISI Web of Science)[22], qui montrera notamment que les modèles harmoniques, qui sont d'une grande robustesse et peuvent utiliser des représentations réduites à différentes échelles (allant du tout atome à un pseudo-atome pour dix résidus)[23, 24, 25], permettent également de rendre compte de la dynamique fonctionnelle des protéines[26, 27, 28], de leur structuration en domaines[29, 30, 31] ou encore de changements conformationnels de grande amplitude[20, 32, 33].

1.1.3 Un modèle gros-grain pour les études de docking : Le modèle de Zacharias

Si la plupart des modèles gros-grain sont exploités pour l'étude des interactions intra-protéiques lors des phénomènes de repliement ou de transition conformationnelle, des représentations simplifiées ont également été développées pour modéliser les interaction inter-protéiques. C'est le cas du modèle réduit de M. Zacharias[34] qui a été mis au point pour des simulations de « docking », ou amarrage, prenant en compte l'effet du mouvement des chaînes latérales lors du processus de reconnaissance protéique via l'utilisation d'une banque de rotamères. Dans cette représentation, chaque acide aminé comprend un pseudo-atome centré sur son carbone α, et un à deux pseudo-atomes qui servent à représenter la chaîne latérale (à l'exception des glycines qui sont modélisées par un unique pseudo-atome). Ce modèle qui traite les protéines comme des corps rigides ne comprend pas de potentiel d'interaction intramoléculaire, tandis que les particules appartenant à deux protéines distinctes interagissent alors selon un potentiel intermoléculaire simplifié de la forme :

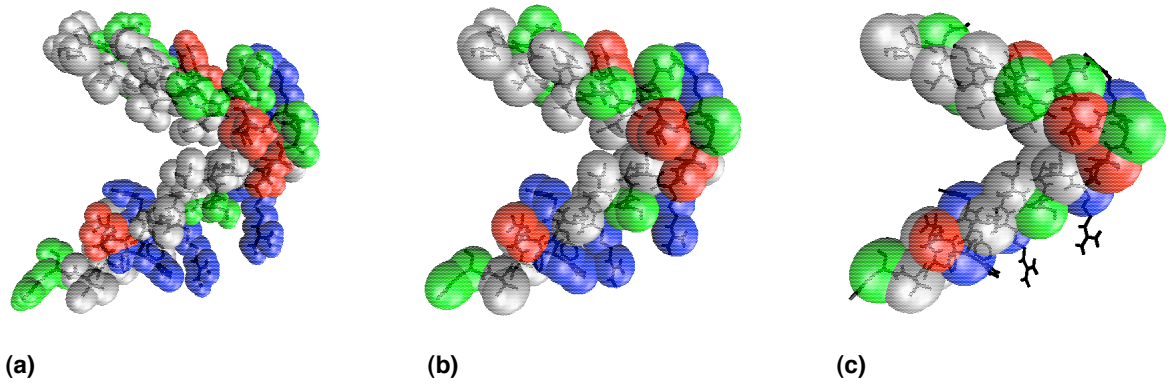


FIG. 1.5 – *Fragment de chaîne protéique en trois réprésentations : (a) tout-atome, (b) modèle de Zacharias, (c) modèle à un pseudo-atome par résidu.*

$$E\left(r_{ij}\right) = \left(\frac{B_{ij}}{r_{ij}^{8}} - \frac{C_{ij}}{r_{ij}^{6}}\right) + \frac{q_i q_j}{\epsilon\left(r_{ij}\right) r_{ij}}, \text{ où } \epsilon\left(r_{ij}\right) = 15 r_{ij} \tag{1.2}$$

Les paramètres B_{ij} et C_{ij} rendent compte de la taille et des propritétés physico-chimiques des résidus auxquels appartiennent les pseudo-atomes i et j dans un terme d'interaction de type Lennard-Jones, tandis que q_i et q_j indiquent les charges des pseudo-atomes dans le terme électrostatique.

Par rapport aux modèles classiques ne comprenant qu'un seul pseudo-atome par résidu, cette représentation réduite permet notamment de rendre correctement compte de l'occupation de l'espace par les chaînes latérale, tout en conservant des temps de calculs réduits, voir la figure 1.5. C'est ce modèle, qui a été intégré avec succès à plusieurs programmes de docking[35, 36, 37, 38, 39], que j'ai choisi pour servir de base à mon travail sur la mécanique des protéines et les interactions protéiques.

Chapitre 2

Étude des propriétés mécaniques des protéines

Si les structures tridimensionnelles d'un très grand nombre de protéines sont désormais résolues ou en voie de l'être, ces informations structurales sont néanmoins insuffisantes pour comprendre les propriétés mécaniques et dynamiques des protéines. Or il a été démontré que les mouvements internes des macromolécules sont indispensables à leur bon fonctionnement biologique[40]. Les simulations de Dynamique Moléculaire permettent de faire le lien entre les propriétés structurales et dynamiques des protéines, mais elles restent cependant limitées à des échelles de temps courtes (de l'ordre de la centaine de nanosecondes) et ne peuvent donc fournir des informations que sur des petits mouvements locaux ou d'échelle intermédiaire. L'Analyse des Modes Normaux d'une protéine permet quant à elle d'obtenir des informations sur les mouvements collectifs au sein de la molécule, notamment sur les déplacements des différents domaines protéiques[19, 2].

2.1 Probing Proteins Heterogeneity : Le programme ProPHet

En complément des approches théoriques citées plus haut, j'ai donc souhaité développer une approche originale permettant d'obtenir des informations concernant les propriétés mécaniques de la protéine à l'échelle du résidu. Je me suis tout particulièrement penchée sur le cas des résidus situés au niveau du site actif et dont les études expérimentales ont montré qu'ils présentent une flexibilité réduite par rapport au reste de la protéine[41]. C'est ainsi que, lors de mon séjour post-doctoral au LBT avec R. Lavery, j'ai mis au point le programme ProPHet (Probing Protein Heterogeneity) qui associe une représentation réduite des protéines en réseau

élastique [24, 34] et un algorithme de dynamique brownienne[42] afin de reproduire les fluctuations dynamiques d'une protéine autour de sa conformation native.

ProPHet reprend le modèle gros-grain des protéines développé par M. Zacharias et détaillé plus haut, auquel il associe un potentiel harmonique. Les « noeuds » du réseau ainsi formé qui sont situés à moins de 9 Å dans la structure initiale de la protéine étudiée interagissent via un ressort présentant une constante de force γ de 0.6 kcal mol^{-1} Å^{-2}. Cette valeur est légèrement inférieure aux valeurs types trouvées dans la littérature, et qui sont de l'ordre de 1.0 kcal mol^{-1} Å^{-2}[24, 25, 28], afin de compenser la densité plus importante du modèle de Zacharias par rapport aux représentations ne comportant qu'un pseudo-atome par résidu. Les mouvements internes de l'objet élastique ainsi créé sont ensuite modélisés via un algorithme de dynamique brownienne où le déplacement de chaque particule du système suit l'équations d'Ermak et McCammon[42],

$$\boldsymbol{r}_i = \boldsymbol{r}_i^0 + \sum_j \frac{\boldsymbol{D}_{ij}^0 \boldsymbol{F}_j^0}{kT} \Delta t + \boldsymbol{R}_i \left(\Delta t\right), \tag{2.1}$$

où $\boldsymbol{r}_i$ et $\boldsymbol{r}_i^0$ correspondent aux vecteurs positions de la particule i avant et après un pas de temps Δt. $\boldsymbol{D}_{ij}$ est un tenseur de diffusion dépendant de la configuration du système et $\boldsymbol{F}_i$ correspond aux forces appliquées sur la particule i. $\boldsymbol{R}_i \left(\Delta t\right)$ est un déplacement aléatoire qui présente une distribution gaussienne, une valeur moyenne nulle, et dont la covariance est définie par :

$$\left\langle \boldsymbol{R}_i \left(\Delta t\right) \boldsymbol{R}_j \left(\Delta t\right) \right\rangle = 2\boldsymbol{D}_{ij}^0 \Delta t. \tag{2.2}$$

Afin de décrire correctement la dynamique du système, les interactions hydrodynamiques entre particules doivent être prises en compte et sont intégrées au problème via le tenseur de diffusion $\boldsymbol{D}_{ij}$ qui reprend la formule de Rotne-Prager[43]. La combinaison d'une représentations réduite des protéines et d'une modèle de solvant implicite nous permet d'utiliser un pas de temps $\Delta t = 10 fs$ nettement supérieur à celui employé dans une dynamique moléculaire tout atome [44].

Les simulations sont ensuite analysées en terme de fluctuations des distances moyennes entre particules. L'inverse de ces fluctuations nous donne alors une « constante de force » k_i, qui sera d'autant plus importante que le système étudié est difficilement déformable au niveau du pseudo-atome considéré, selon la formule :

$$k_i = \frac{3k_B T}{\left\langle \left(d_i - \left\langle d_i \right\rangle\right)^2 \right\rangle}, \tag{2.3}$$

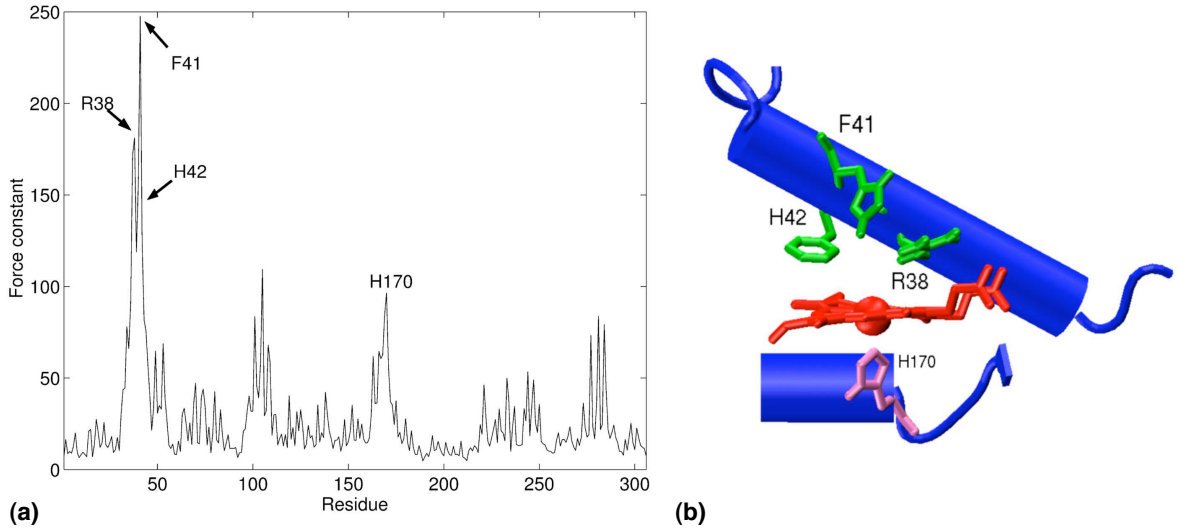


FIG. 2.1 – *(a) Profil de rigidité de Horseradish Peroxidase c (HRPC), les résidus annotés sont situés au niveau du site actif. (b) Vue agrandie du site actif de HRPC comprenant les résidus rigides identifiés sur la figure précédente et le cofacteur hème en rouge.*

où $\langle\rangle$ représente une valeur moyenne sur l'ensemble de la simulation et $d_i = \langle d_{ij} \rangle_{j*}$ est la distance moyenne entre la particule i et les autres particules j (la somme sur l'ensemble réduit $j*$ exclut les pseudo-atomes appartenant au même résidu que i). Finalement, la constante de force d'un acide aminé est simplement obtenue en faisant la moyenne des constantes de force de ses pseudo-atomes constitutifs et les propriétés mécaniques d'une protéine peuvent alors être présentées sous la forme d'un profil de rigidité affichant les constantes de force de toute sa séquence d'acide aminés (voir par exemple la figure 2.1.a).

Comparée aux méthodes développées précédemment au LBT pour étudier la mécanique des protéines (et qui nécessitaient plusieurs calculs de minimisation de l'énergie par résidu, [31]), l'approche de ProPHet permet d'obtenir un profil de rigidité complet en une seule simulation, ce qui représente un gain très important en temps de calcul et va rendre possible l'étude de systèmes de grande taille.

2.2 Utilisation autonome de ProPHet

2.2.1 Prédiction des résidus catalytiques au sein des protéines

Protéines à hème

ProPHet a tout d'abord été exploité pour étudier un ensemble de protéines comportant un ou deux cofacteurs de type hème[45]. Ces premiers travaux ont mis en évidence la rigidité accrue des résidus situés au niveau du site actif des protéines. Ainsi, sur la figure 2.1, les pics de rigidité de horseradish peroxidase correspondent aux résidus catalytiques Arg38, Phe41, His 42 et His170. La comparaison des profils de rigidités obtenus à partir de différentes structures d'une même protéine reflète également son mode de fonctionnement biologique. La cytochrome c peroxydase par exemple, présente dans sa forme inactive un groupement heme hexacoordiné par deux histidines His55 et His71, voir la figure 2.2.b. Lors de son passage à la forme active, qui est pentacoordinée, on observe une libération de l'histidine distale His71, ce qui va permettre l'accès du substrat au site actif [46]. La dissymétrie entre les deux histidines distale His71 et proximale His55 apparaît dans le profile de rigidité de la protéine dans sa forme inactive (figure 2.2.a, ligne supérieure), où His55 présente un pic de rigidité marqué, alors que His71 est elle particulièrement flexible. Cette flexibilité inhabituelle pour un résidu appartenant au site catalytique est néanmoins nécessaire pour permettre le passage à la forme active de la protéine via un mouvement de grande ampleur de His71. Si l'on regarde les variations mécaniques induites par le changement conformationnel lors de la transition forme inactive → forme active sur la figue 2.2.c, on observe une rigidification accrue de l'ensemble du site catalytique à l'exception des résidus His71 et Phe93 qui vont s'écarter afin de permettre la fixation d'un substrat.

L'analyse des propriétés mécaniques des protéines comportant un domaine structural de type cytochrome-c permet également de détecter un autre type de résidus rigides qui ne jouent pas un rôle fonctionnel mais structural dans la protéine. Ces résidus apparaissent dans les profils de rigidité sous la forme de deux doublets de pics (voir la figure 2.2.a, ligne inférieure) et occupent les positions *(i,i+3)* et *(j,j+4)* dans deux hélices α orthogonales, les positions *(i+3)* et *(j+4)* correspondant à des acides aminés possédant des chaînes aromatiques (figure 2.2.d). Il s'avère que ces quatre résidus, dont les chaînes latérales sont en forte interaction, forment ce que l'on appelle le noyau de repliement du domaine cytochrome-c, soit un ensemble d'acides aminés fortement conservés dans cette famille protéique et qui va jouer un rôle clé lors du processus de repliement[47, 48].

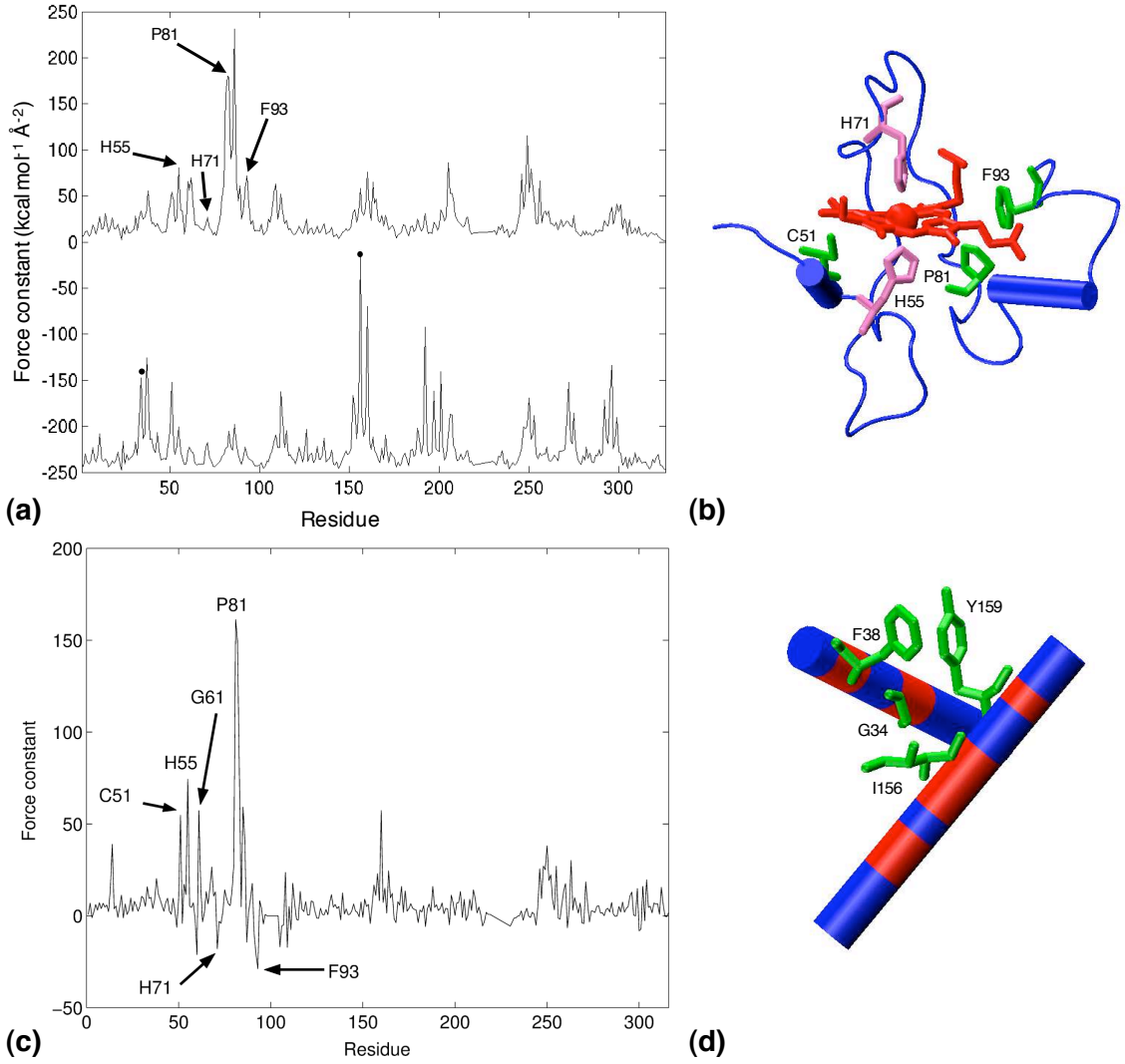


FIG. 2.2 – *(a) Profil de rigidité de la cytochrome-c peroxydase (CCP) dans sa forme inactive, analyse de la structure complète (ligne supérieure), ou analyse en séparant les domaines structuraux[45] (ligne inférieure avec un décalage de -250 kcal mol^{-1} Å^{-2}. Les points noirs sur la ligne inférieure signalent les deux doublets de pics du noyau de repliement. (b) Site catalytique de la CCP (c) Variation du profil de rigidité de la CCP lors du passage de la forme inactive à la forme active (d) Noyau de repliement de la CCP, les deux doublets de résidus sont marqués par des points noirs sur la ligne inférieure de la figure 2.2a*

Travail sur une base protéique élargie

Suite à cette première étude portant sur une famille protéique restreinte, nous avons effectué une analyse systématique des propriétés mécaniques d'une centaine d'enzymes lors du stage de M1 au LBT d'É. Laforet [49]. les données statistiques ainsi obtenues mettent bien en évidence

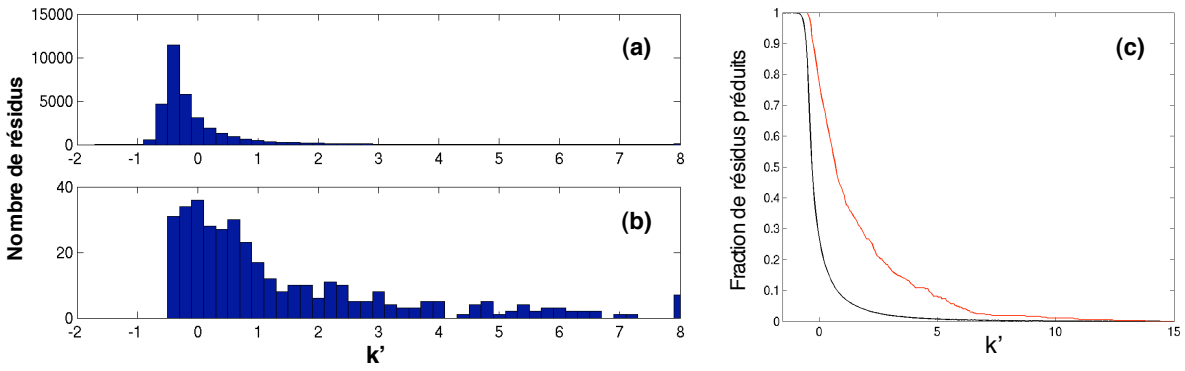


FIG. 2.3 – *Distribution des constantes de force normalisées k' sur la base protéique : (a) Tous les résidus, (b) Résidus catalytiques. (c) Proportion de résidus détectés en fonction de k' pour l'ensemble des résidus (ligne noire) ou les résidus catalytiques (ligne rouge).*

les propriétés mécaniques spécifiques des résidus catalytique au sein des enzymes étudiées. L'amplitude des variations des constantes de force k observées au sein d'une protéine dépendant directement de la taille de celle-ci, afin de pouvoir comparer les résultats de l'ensemble de la base de données enzymatique[28] nous avons défini une constante de force normalisée k' selon la formule suivante :

$$k' = \frac{k - \langle k \rangle}{\sigma\left(k\right)} \tag{2.4}$$

où $\langle k \rangle$ est la valeur moyenne des constantes de forces des résidus d'une protéine donnée et $\sigma\left(k\right)$ leur écart type. Une valeur positive de k' signale donc un résidu présentant une rigidité accrue par rapport à la moyenne protéique. Les histogrammes des figures 2.3.a et b. montrent un fort déplacement de la distribution des constantes de forces normalisées vers les valeurs positives dans le cas des résidus catalytiques par rapport aux autres résidus ; ce qui permet d'envisager une utilisation de k' comme critère pour la prédiction du site catalytique au sein d'une protéine. Ainsi, sur la base enzymatique étudiée (qui regroupe 98 protéines totalisant plus de 33000 résidus dont 370 sont répertoriés expérimentalement comme résidus catalytiques), 28% des résidus et 78% des résidus catalytiques présentent une valeur de k' positive (figure 2.3.c).

2.2.2 Réponse d'une protéine soumise à une contrainte mécanique externe

Les dernières années ont vu le développement d'un grand nombre de techniques, telles que les pinces optiques ou magnétiques[50, 51, 52], ou le microscope à force électronique[53, 54, 55],

qui ont rendu possible la manipulation de systèmes biologiques à l'échelle moléculaire[56]. Ces expériences se sont focalisées dans un premier temps sur la réponse mécanique d'une protéine soumise à une force externe, en suivant par exemple la séquence d'événements de dépliement de la chaîne protéique observés lors de l'extension de celle-ci[53]. Plus récemment de nouvelles manipulations ont permis d'observer la réponse fonctionnelle d'une enzyme sous contrainte mécanique, ouvrant ainsi la voie au vaste champ de la mécanoenzymatique[57, 58].

C'est dans cette perspective que ProPHet a été modifié afin de permettre l'introduction d'une contrainte externe sur le système protéique étudié. Cette implémentation nous a alors permis de modéliser simplement deux expériences récemment mises en place et d'étudier ainsi la réponse mécanique ou fonctionnelle d'une protéine soumise à des tensions selon différentes directions.

Réponse mécanique : Etude de la Green Fluorescent Protein

Je me suis tout d'abord penchée sur les travaux expérimentaux réalisés par Dietz *et al.* sur la Green Fluorescent Protein (GFP)[59]. Alors que les manipulations d'extension de molécule

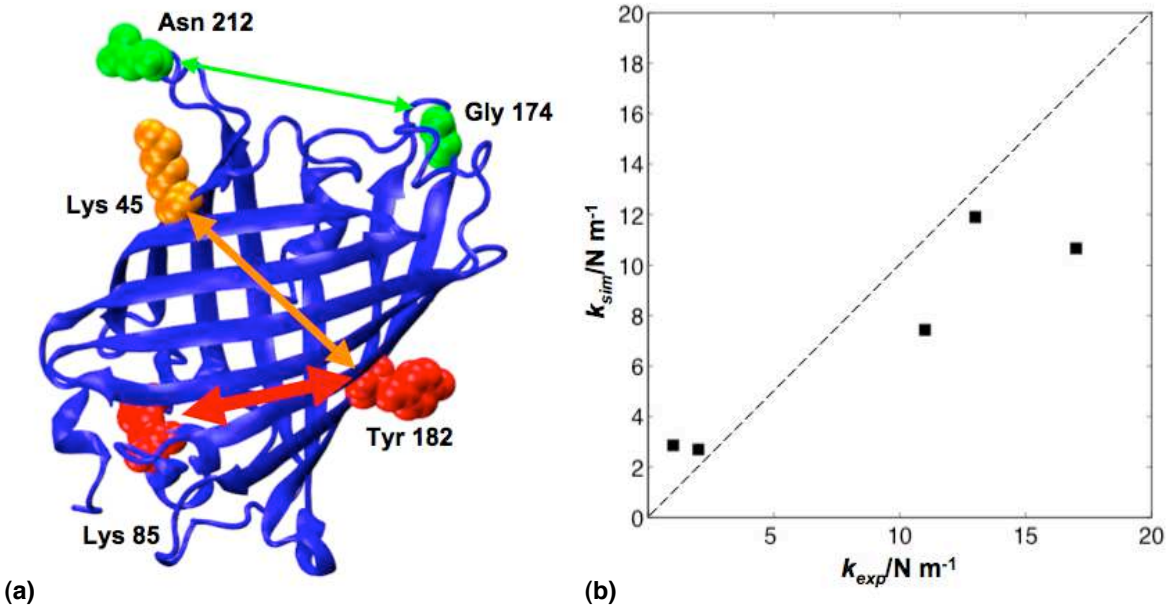


FIG. 2.4 – *(a) Trois directions de déformation types pour GFP entraînant une réponse mécanique faible (en vert), moyenne (en orange) ou importante (en rouge), (b) Comparaison des constantes de force directionnelles obtenue expérimentalement (abcisse) et via les simulations avec ProPHet (ordonnée)*

unique se font généralement le long de l'axe N-C-terminal de la protéine, cette équipe a réussi à déformer la GFP selon trois axes différents, mettant ainsi en évidence la forte anisotropie de la GFP sous contrainte et la relation entre la structure secondaire de la protéine et sa résistance mécanique, voir la figure 2.4.a.

Dans la version modifiée de ProPHet, la contrainte externe est modélisée par le biais d'une force constante appliquée entre deux résidus et les constantes de force directionnelles sont calculées à partir la variation de la distance moyenne observée entre les deux points d'ancrage lors de simulations réalisées avec ou sans contrainte externe[60]. On obtient au final une excellente corrélation (0.94) entre les données expérimentales et théoriques, voir la figure 2.4.b. Une étude plus systématique de la déformation de la protéine selon plus de 300 directions est ensuite venue souligner le lien entre l'agencement des éléments de structure secondaire de la molécule et les constantes de force directionnelles obtenues lors de sa déformation. Ainsi, dans la structure en tonneau β de la GFP les axes « rigides » sont typiquement orientés parallèlement aux feuillets β de la protéine (axe rouge sur la figure 2.4.a), tandis que les résidus qui ne sont pas directement reliés par un élément de structure secondaire formeront des axes de déformation plus souples (axe orange) ou même très flexibles dans le cas des résidus situés à l'extrémités d'une boucle (axe vert). Les constantes de force directionnelles obtenues vont alors de 1.5 Nm^{-1} à 60 Nm^{-1} avec une valeur moyenne autour 15 Nm^{-1}.

Des calculs supplémentaires ont ensuite été réalisés sur trois autres protéines ayant aussi fait l'objet de manipulations selon différentes directions[61, 62, 63] et nos résultats reproduisent systématiquement les propriétés d'anisotropie mécanique observées expérimentalement.

Réponse fonctionnelle : Etude de Guanylate Kinase

On peut également modéliser l'application d'une force extérieure sur une protéine en introduisant dans le modèle en réseau élastique employé dans ProPHet un ressort supplémentaire, dit ressort de contrainte, positionné entre les points d'application de cette force et dont les caractéristiques (raideur et longueur à vide) vont être choisies à partir des données expérimentales[65]. C'est cette méthode qui a été utilisée pour représenter le dispositif de « sonde allostérique », développé dans l'équipe de biophysique de G. Zocchi à UCLA, tel qu'il a été mis en place sur Guanylate Kinase (GK)[66, 67, 68]. Cette enzyme catalyse le transfert d'un groupement phosphate de l'ATP vers GMP, la réaction enzymatique étant accompagnée d'un mouvement d'ouverture-fermeture autour du site catalytique, voir la figure 2.5.b. Lors de leur expérience, Zocchi *et al.* ont étudié comment la direction de la contrainte appliquée à la protéine pouvait influer sur son activité enzymatique en mesurant les constantes d'affinité pour les différents ligands de la protéine, ainsi que l'évolution du taux catalytique en fonction de la contrainte exercée.

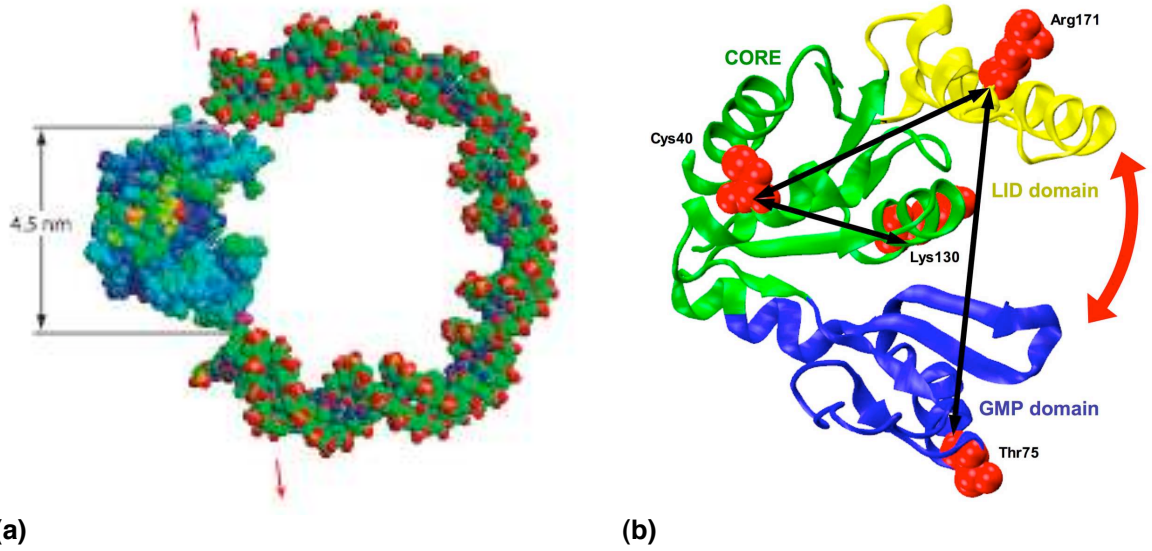


FIG. 2.5 – *(a) Chimère protéine-ADN permettant d'étudier la réponse fonctionnelle de GK soumise à une contrainte mécanique[64](b)Trois directions de déformations testées expérimentalement pour GK. La flêche rouge indique le mouvement d'ouverture/fermeture de la protéine qui a lieu durant la réaction enzymatique.*

Dans le cadre du **projet ANR FonFloN**, qui vise à faire le lien entre **fonction** enzymatique et **fluctuations** structurales dans les protéines, et en collaboration avec M. Baaden et O. Delalande, j'ai effectué des calculs sur GK à l'aide de ProPHet, qui nous ont permis d'apporter une explication atomistique aux observations expérimentales[69]. En effet nous avons pu montrer que la première direction de contrainte testée (Thr75/Arg171) entraîne un flexibilisation du site de fixation du ligand GMP, ce qui va induire une baisse de l'affinité observée pour ce ligand. La deuxième direction testée (Cys40/Arg171) va quant à elle occasionner des modifications importantes du premier mode normal de vibration de la protéine. Or ce premier mode correspond à la transition conformationnelle observée pendant la réaction enzymatique, sa perturbation peut donc être mise en rapport avec la baisse du taux d'activité catalytique obtenue lors des manipulations.

En complément des travaux expérimentaux, nous avons également effectué une étude plus systématique de la déformation de la protéine en testant plus de deux cents points d'ancrage potentiels au niveau de la surface moléculaire. Nous avons ainsi mis en évidence une nouvelle direction de perturbation (Asp65/Leu122), qui devrait entraîner simultanément des modification de l'affinité pour GMP *et* une baisse du taux catalytique. Nos résultats ont été communiqués à l'équipe de G. Zocchi pour une éventuelle confirmation expérimentale.

Plus généralement, ces travaux de « mécanoenzymatique » montrent qu'il est possible de rendre compte, à l'aide de modèles simples, des expériences réalisées sur les protéines, et comment leurs propriétés mécaniques et dynamiques peuvent contrôler leur activité biologique.

2.3 Utilisation en complément d'autres méthodes

2.3.1 Association avec des méthodes expérimentales, le centre réactionnel de R. Sphaeroides

Dans le cadre d'une collaboration avec l'équipe de biophysique du Pr. P. Sebban du Laboratoire de Chimie Physique de l'Université d'Orsay, j'ai également utilisé ProPHet pour étudier la dynamique interne d'une grosse protéine membranaire, le centre réactionnel (CR) de *Rhodobacter Sphaeroides*[70]. Cette protéine, qui comprend trois sous-unités (H, L et M) et plus de 800 acides aminés, est un ancêtre du photosystème II que l'on retrouve dans les organismes

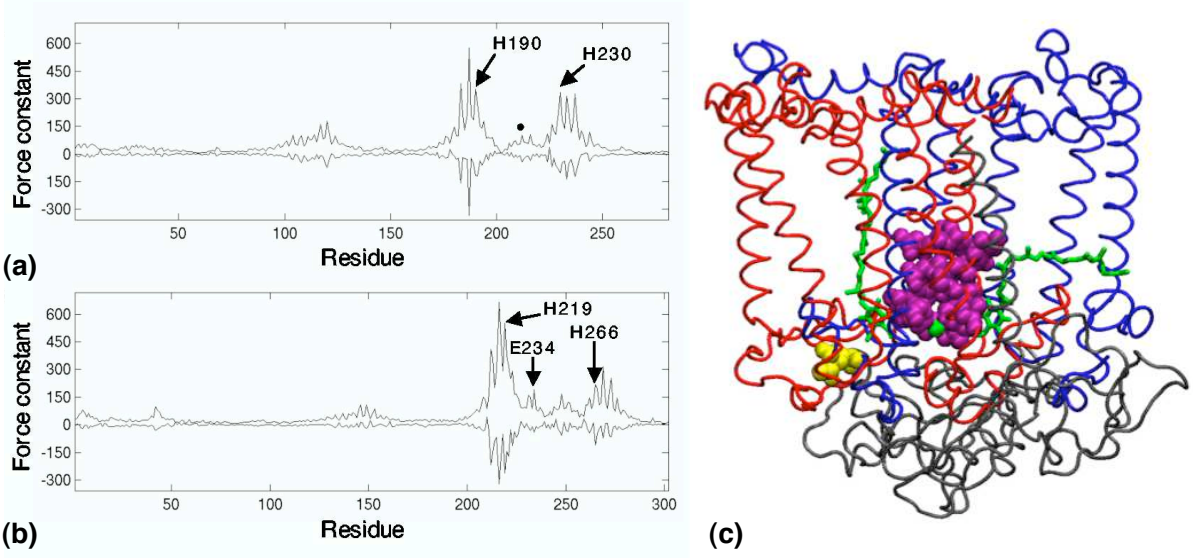


FIG. 2.6 – *Profils de rigidité du CR de R. Sphaeroides, la ligne supérieure indique les constantes de forces des résidus dans la protéine native, tandis que la ligne inférieure indique les variations de flexibilité quand on passe au mutant AA, les résidus mutés sont signalés par un point noir, (a) chaîne L, (b) chaîne M. (c) Structure du mutant AA du CR de R. Sphaeroides avec les chaînes L (bleue), M (rouge) et H (noir). Les quinones Q_A et Q_B, ainsi que l'atome de fer hors-hème impliqué dans le transfert d'électron, sont indiqués en vert. Les résidus mutés L212 et L213 sont en jaune et les résidus dont les propriétés mécaniques sont les plus perturbées par la mutation sont en violet.*

photosynthétiques supérieurs et convertit la lumière en énergie chimique via une chaîne de transferts couplés d'électrons et de protons [71, 72, 73]. Parmi ces transferts on a le passage de deux électrons de la quinone Q_A vers la quinone Q_B via un atome de fer (hors-hème) pentacoordiné, suivi d'une double protonation afin de former une dihydroquinone Q_BH_2. Dans le mutant L212Glu/L213Asp → Ala/Ala (AA), les transferts de proton vers Q_B sont réduits par un facteur 10^3, entravant ainsi l'activité biologique du CR [74, 75]. Des expériences de diffusion de neutron sur le CR natif et le mutant AA ont mis en évidence une flexibilité accrue de celui-ci par rapport à la protéine native au delà de la température de transition dynamique T_d[40], bien que les études cristallographiques ne permettent pas d'observer de différences structurales notables dans le squelette protéique liées à la mutation [76]. Les profils de rigidité produits par ProPHet à partir de ces structures cristallographiques, par contre, montrent bien une variation importante des propriétés mécaniques du mutant par rapport à la protéine native, avec notamment une flexibilisation accrue des résidus situés au cœur du CR, et tout particulièrement les ligands de l'atome de fer hors-hème, voir les figures2.6.a et b.

Suite à ces travaux, j'ai été contactée par G. Venturoli, de l'Université de Bologne, qui travaille également sur ce système, notamment en étudiant la dynamique interne de la protéine lorsque celle-ci est insérée dans une matrice de trehalose [77]. À sa demande, j'ai réalisé des calculs sur le centre réactionnel dans sa structure native (wtCR) et dans une forme où un ligand caroténoide a été enlevé (R26CR). Cette fois encore, les profils de rigidités obtenus mettent en évidence la flexibilité accrue de la variété R26, notamment autour de l'atome de fer situé au cœur de la protéine. Ces résultats concordent avec les données cinétiques obtenues par spectroscopie qui montrent que la dynamique interne de la protéine est plus inhibée pour la forme wt que pour la variété R26 [78].

Dans les deux cas, ces résultats soulignent l'importance des propriétés mécaniques des protéines pour leur bon fonctionnement biologique et leur dynamique interne. En complément des expériences de diffusion de neutron et de spectroscopie, les profils de rigidité permettent de localiser précisément les variations dynamiques induites par des mutations locales et montrent que celles-ci peuvent voir des effets à longue portée. En effet, dans le cas du mutant AA, les zones qui subissent la plus forte baisse de rigidité sont situés en moyenne à 15Å des résidus mutés, voir la figure 2.6.c, et pour la variété R26, les résidus fortement flexibilisés se situent à au moins 19Å du caroténoide.

2.3.2 Association avec des simulations tout atomes

Diffusion de petits ligands dans la neuroglobine

Membre de la famille des globines découverte récemment [79], la neuroglobine (Ngb) est exprimée en petite quantité dans le cerveau dans des conditions d'hypoxie et son rôle biologique reste encore à élucider[80]. En effet, les travaux expérimentaux concernant cette protéine ont mis en évidence l'hexacoordination de l'atome de fer en l'absence de ligand, ce qui entraîne une faible affinité pour l'oxygène, excluant ainsi une fonction biologique de type transport/stockage de O_2 pour Ngb à la différences des autres globines. Les études structurales montrent que la transition d'une forme hexa- vers une forme penta-coordinée préliminaire à la fixation d'un

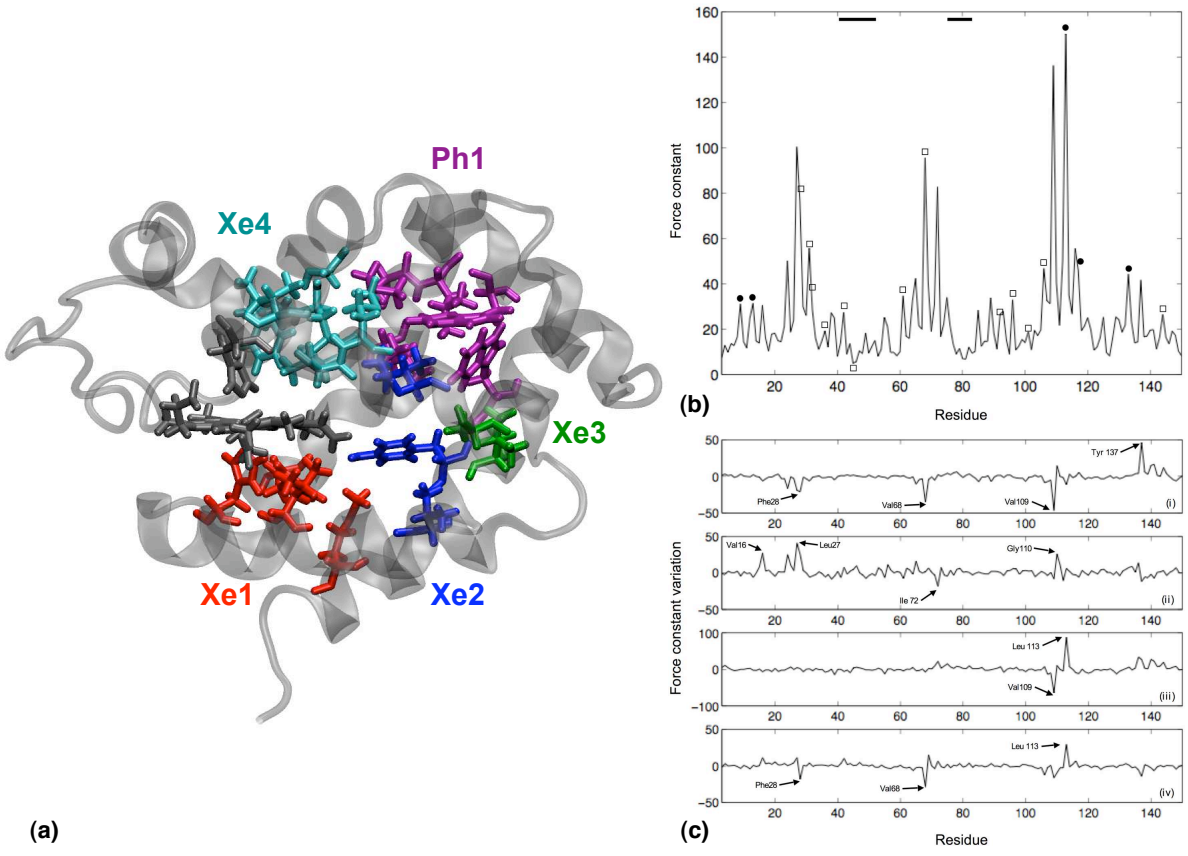


FIG. 2.7 – *(a) Réseau de cavités internes de la neuroglobine humaine (hNgb) (b) Profil de rigidité de hNgb dans sa forme hexacoordinée et sans disulfure. Les résidus conservés pour l'ensemble des globines sont marqués soit par des carrés blancs (résidus lié à l'hème) ou des ronds noirs (noyau de repliement [48]). (c) Variations du profil de rigidité lors des changements d'état (coordination ou redox) de hNgb. Les résidus annotés sont situés soit à la frontière entre deux cavités internes, soit entre une cavité interne et l'extérieur de la protéine.*

ligand sur le fer se fait non pas suite à une rotation de l'histidine distale His64, mais plutôt par le biais d'un « glissement » de l'hème au sein de la protéine, ce qui va entraîner des réarrangements considérables de son réseau de cavités internes[81]. Une autre spécificité de Ngb dans sa forme humaine est la présence de deux cystéines (Cys46 et Cys55) dans la boucle flexible CD qui sont susceptibles de former un pont disulfure. Les travaux expérimentaux ont montré que le changement d'état d'oxydation de la protéine influe sur l'affinité du fer pour l'histidine distale [82]. La réduction des deux cystéines favorisant la forme hexacoordinée du fer, la Ngb pourrait libérer de l'oxygène en condition d'hypoxie et jouer ainsi un rôle de signalement qui permettrait le déclenchement d'un processus de protection des cellules [83].

Dans cette perspective et dans le cadre de la première année de thèse d'Anthony Bochaut, nous avons étudié la diffusion de petits ligands (CO, NO et O_2) au sein du réseau de cavités de la Ngb humaine[84], notamment afin de mieux comprendre comment l'état redox des cystéines de la boucle CD pouvait influer sur celle-ci. Lors de ce travail, où plusieurs techniques de simulations (dynamique moléculaire, métadynamique) ont été exploitées, les calculs réalisés avec ProPHet sur différentes formes de Ngb (penta- et hexa-coordinées, avec ou sans pont disulfure) ont permis de mettre en évidence les propriétés mécaniques spécifiques des résidus situés au niveau des frontières entre cavités internes, voir la figure 2.7. Les transitions conformationnelles dues à des changements de coordination ou de l'état redox de la protéine donnent en effet lieu à une réorganisation importante du réseau de cavités de la Ngb qui s'accompagne de variations du profil de rigidité localisées spécifiquement au niveau des résidus frontière. Ces variations semblent jouer un rôle dans le contrôle du passage du ligand d'une cavité à une autre ou d'une cavité interne vers le solvant et ont pu être reliées aux parcours de diffusion des petits ligands observés lors des calculs par métadynamique.

Mécanique des résidus frontière dans les globines

Dans leur travail sur la myoglobine[85], Scorciapino *et al.* ont mis en évidence le rôle clé de certains résidus localisés à la frontière entre deux cavités internes pour la régulation du passage d'un petit ligand. Il s'avère que ces résidus clés occupent des positions analogues aux résidus frontière identifiés lors de notre premier travail sur la neuroglobine, et nous avons donc entrepris d'étudier un jeu de globines élargi comprenant six chaînes (myoglobine, neuroglobine, cytoglobine, hémoglobine tronquée, et les chaînes α et β de l'hémoglobine humaine), afin de voir dans quelle mesure les propriétés mécaniques des résidus frontière sont conservés au sein de cette famille protéique[86]. Pour chacune des protéines de notre étude, nous avons réalisé des simulations de dynamique moléculaire classiques afin de générer un jeu de cinq structures représentatives, dont les profils de rigidité ont ensuite été établis avec ProPHet. La comparaison

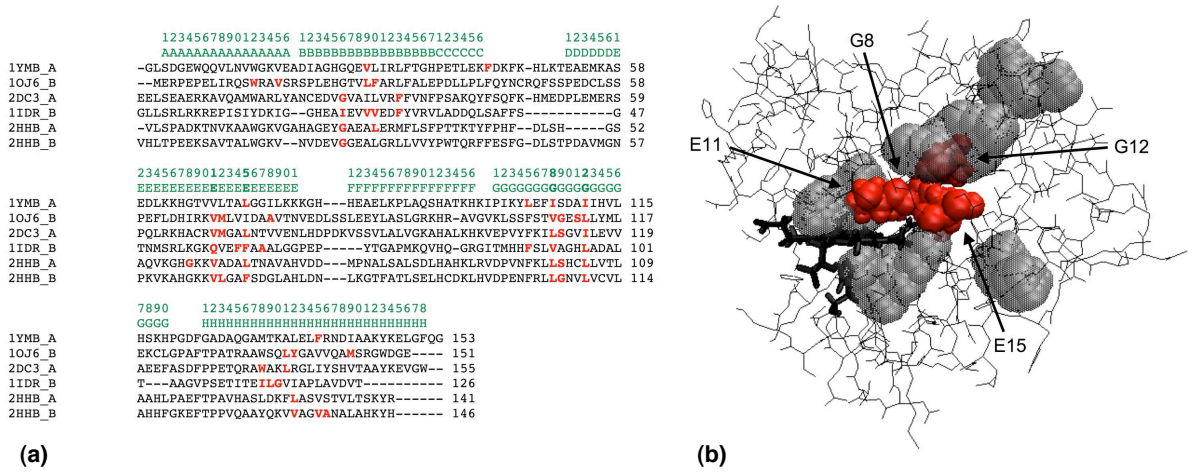


FIG. 2.8 – *(a) Alignement des six séquences de globines, les résidus mécaniquement sensibles sont marqués en rouge (b) Noyau mécanique, en rouge, au cœur du réseau de cavités internes de la myoglobine.*

deux à deux des cinq profils de chaque globine met évidence un nombre restreint (de l'ordre de la dizaine) de résidus « mécaniquement sensibles ». Ces résidus sont systématiquement situés en bordure d'une cavité interne, et l'alignement des six séquences protéiques montre que leurs positions sont extrêmement bien conservées. On voit notamment apparaître au cœur de la protéine un « noyau mécanique » correspondant aux résidus occupant les positions E11, E15, G8 et G12 dans le repliement caractéristique des globines, voir la figure 2.8. Les résidus formant ce noyau apparaissent comme étant mécaniquement sensibles dans les six chaînes protéiques de notre étude et chacun d'entre eux a déjà déjà fait l'objet de travaux montrant son importance dans le contrôle de la migration d'un ligand au sein du réseau de cavités internes pour au moins une globine. Notre approche permet donc, en combinant simplement simulations tout atome et gros-grain, de mettre en évidence des résidus clés pour le fonctionnement de toute une famille protéique et pourrait être appliquée aux nombreuses protéines globulaire présentant des canaux de diffusion internes.

2.4 Conclusion

Le développement et l'utilisation de ProPHet pour l'étude d'un grand nombre de systèmes ont montré comment la diversité des structures protéiques entraîne des comportements mécaniques variés, qui sont eux mêmes à l'origine de nombreux processus biologiques tels que le transport de ligands, les transferts d'électrons ou encore la catalyse enzymatique. L'exploitation d'un modèle gros-grain pour les protéines permet donc de souligner comment les propriétés

mécaniques représentent un maillon essentiel dans la relation structure-fonction qui, plus de cinquante ans après la résolution des premières structures protéiques[87, 88], reste encore au cœur de notre façon d'appréhender la biologie moléculaire.

Chapitre 3

Interactions protéiques

3.1 Molecular Association via Cross-Docking : Le programme MAXDo

3.1.1 Introduction

Les interactions protéiques jouent un rôle central dans l'exécution, la coordination et la régulation des activités biologiques, ce qui fait de la compréhension de l'*interactome* des organismes un élément clé de notre approche du vivant[89, 90, 91, 92]. De nombreuses techniques expérimentales telles que les méthodes de double hybride ou TAP (Tandem Affinity Purification)[93] ont permis de cartographier les interactions protéiques d'un certain nombre d'organismes comme la levure[94], *E. Coli*[95] ou même l'homme[96]. Néanmoins ces méthodes restent coûteuses à mettre en place et génèrent un grand nombre de faux positifs et négatifs qui réduisent considérablement leur précision[97, 98]. Il existe également des méthodes *in silico* basées sur l'analyse des séquences protéiques[99], mais ces approches ne fournissent pas d'information au niveau atomique sur la conformation des complexes où sur les interactions mises en jeu au sein de ceux-ci.

La modélisation moléculaire est une alternative pour la prédictions des interactions protéiques qui permet également d'obtenir des informations concernant les modes d'interactions des protéines sur le plan structural. Cependant celle-ci s'est longtemps restreinte au seul problème de l'amarrage protéique (ou docking), qui consiste à prédire la conformation d'un complexe à partir des structures isolées de ses composants[100]. En effet, la prédiction des partenaires d'interaction potentiels au sein d'une base de données protéiques, soit le problème du « pre-docking », demeure extrêmement coûteuse sur le plan des temps de calculs mis en jeu et est donc restée longtemps inaccessible.

Ce travail, qui fait à l'origine partie des trois projets sélectionnés dans le cadre du programme DECRYPTHON[1] (mis en place par l'Association Française contre les Myopathies, le CNRS et IBM) pour l'année 2005, entend donc réaliser une étude à grande échelle des interactions protéine-protéine afin de mieux comprendre leur spécificité. Notre objectif est donc de combiner des approches bioinformatiques (mises au point dans le groupe de Génomique Analytique d'A. Carbone) et de modélisation moléculaire (développées au LBT), afin de pouvoir localiser les sites d'interaction à la surface des protéines et identifier les partenaires potentiels d'une protéine donnée au sein d'une base de données comprenant des milliers de structures (du type Protein Data Bank[17]).

3.1.2 Algorithme de docking croisé

Les programmes de docking (ou amarrage) protéique se penchent sur les modalités d'interaction entre deux protéines et ont notamment pour objectif la prédiction de la structure d'un complexe protéique à partir d'informations concernant les partenaires isolés (voir [101, 102] pour des revues récentes sur le sujet). Dans un premier temps nous avons développé un algorithme de docking permettant de rechercher les géométries d'interaction optimales entre deux partenaires protéiques et utilisant la représentation réduite des protéines développpée par M. Zacharias[34].

Pour un couple récepteur (protéine fixe)/ligand (protéine mobile) donné, l'algorithme de docking génère un ensemble de positions de départ pour lesquelles l'énergie d'interaction va être

[1] http ://www.decrypthon.fr

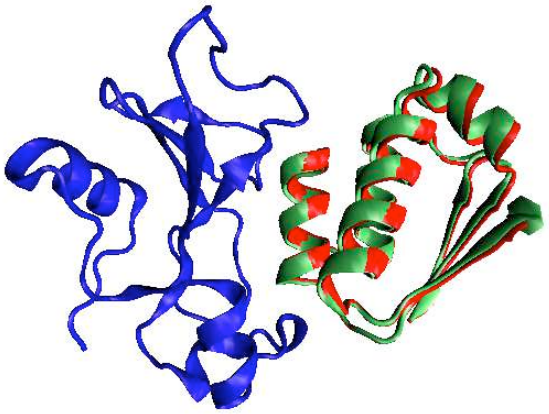


FIG. 3.1 – *Complexe protéique barnase (en bleu)/barstar, avec en rouge la position cristallographique du ligand et en vert sa position prédite par l'algorithme de docking.*

minimisée. Un calcul complet nous permet alors d'établir une carte de la surface énergétique du récepteur pour un ligand donné et notamment de localiser les zones d'interaction favorables à la surface du récepteur. Du fait du très grand nombre de positions de départ du ligand nécessaires pour explorer correctement la surface d'interaction (entre 100000 et 500000 points selon la taille du récepteur), l'algorithme a été spécialement développé pour permettre la mise en place de calculs parallèles et l'exploitation de la grille de calcul universitaire DECRYPTHON, et donc réduire grandement les temps de calcul nécessaires.

Cet algorithme a ensuite été exploité dans le cadre d'une expérience de « Docking Croisé »sur le docking benchmark 2.0, un jeu-test de 84 complexes protéiques [103]. Le processus de docking a été alors appliqué non seulement aux partenaires protéiques connus, mais aussi à la totalité des paires de protéines possibles, qu'il s'agisse de partenaires identifiés expérimentalement ou non. Nous nous sommes donc intéressés pour la première fois à des protéines qui, *en principe*, n'interagissent pas ensemble, ce qui nous permettra d'établir une base de « decoys »(faux positifs) de bonne qualité et qui pourront être exploités dans l'élaboration de potentiels d'interaction. L'ensemble de ces calculs (qui représente plus de 258000 opérations de docking) a été distribué sur le World Community Grid[2](WCG), une grille d'internautes constituée d'ordinateurs personnels et mise en place par IBM.

3.2 Prédiction des partenaires d'interaction protéique

3.2.1 Travail sur une base protéique restreinte

Les premiers calculs effectués sur un ensemble réduit de six complexes protéiques montrent l'efficacité du programme [104]. Dans chaque cas les cartes énergétiques établies par minimisations multiples mettent en évidence un puits de potentiel au niveau de la position cristallographique du ligand par rapport au récepteur. De plus, l'algorithme permet de retrouver pour chacun des complexes une conformation où les atomes du ligand présentent un écart quadratique moyen par rapport à leur position cristallographique inférieur à 3 Å, voir la figure 3.1. Néanmoins, certains « faux » complexes peuvent présenter des énergies d'interactions ou des interfaces comparables à celles des complexes expérimentaux, ce qui souligne le problème que pose actuellement l'identification des partenaires spécifiques au sein d'une large base de données protéique.

Dans une seconde phase d'analyse des données issues des calculs de docking croisé, nous avons mis au point un indice d'association NII (Normalized Interaction Index) qui tient compte **à la fois de l'énergie d'interaction obtenue lors d'un calcul de docking et des résidus**

[2]http ://www.worldcommunitygrid.org

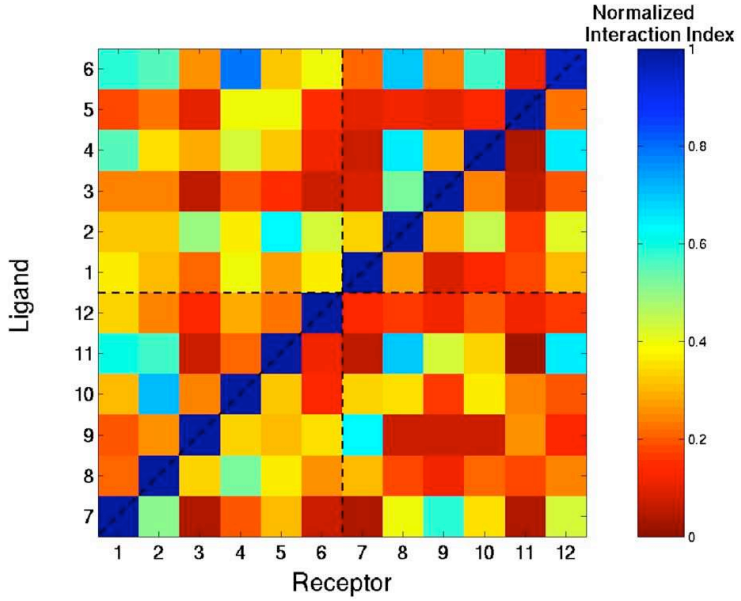


FIG. 3.2 – *Matrice de docking croisé obtenue pour une base réduite de six complexes protéiques, soit douze protéines distinctes. Les protéines ont été ordonnées de manière à ce que les partenaires expérimentaux soit placés sur la diagonale, ceux ci présentent alors systématiquement le meilleur indice d'association.*

présents au niveau de l'interface protéique résultante. Pour l'ensemble des structures générées lors de l'opération de docking, on calcule pour chacun des partenaires (ligand et récepteur) la fraction des résidus appartenant à son interface expérimentale et que l'on retrouve dans l'interface dockée (Fraction of Interface Residues, FIR), ce qui permet d'obtenir un taux global de conservation des résidus d'interface de la forme :

$$FIR = FIR_{rec} \times FIR_{lig} \tag{3.1}$$

Pour un couple protéique P_1-P_2, on définit alors son indice d'interaction (II) comme :

$$II_{P_1P_2} = \max\left(FIR\right)_{P_1P_2} \times E_{tot}\left(\max\left(FIR\right)\right), \tag{3.2}$$

où $\max\left(FIR\right)$ est la valeur maximale du FIR obtenue sur l'ensemble des configurations générées pour le complexe P_1-P_2, et E_{tot} l'énergie d'interaction associée. l'indice II (dont la valeur est négative ou nulle) est ensuite normalisé suivant la formule :

$$NII_{P_1P_2} = \frac{II^2_{P_1P_2}}{\min\left(II_{P_1P_j}\right)_{P_j \in P} \times \min\left(II_{P_jP_2}\right)_{P_j \in P}}, \tag{3.3}$$

afin d'être compris entre 0 et 1. NII est alors d'autant plus important que l'interaction entre les deux protéines est favorable. Pour toutes les protéines de notre base réduite, le partenaire protéique présentant l'indice d'association maximal correspond systématiquement au partenaire expérimental, voir la figure 3.2. Pour la première fois, notre algorithme permet donc de déterminer comment deux protéines vont pouvoir s'associer, mais aussi quelles sont les protéines au sein d'une base de données qui sont susceptibles d'interagir pour former un complexe spécifique.

3.2.2 Analyse du benchmark2.0

L'exploitation de la grille WCG pour effectuer des simulations de docking croisé sur l'ensemble du benchmark 2.0 nous a permis d'évaluer la robustesse de nos prédiction et aussi d'analyser les comportements spécifiques à différentes types de complexes (enzyme/inhibiteur, antigène/anticorps ou autres) en matière de reconnaissance protéique[105].

Naturellement, l'utilisation des structures non liées lors des calculs entraîne une dégradation de la qualité des prédiction des partenaires d'interaction, comme on peut le voir sur la matrice de NII de la figure 3.3a. Néanmoins, lorsque l'on compare la distribution des NII pour l'ensemble des couples protéiques possibles et pour les seuls partenaires expérimentaux, on peut constater un net décalage en faveur des fortes valeur de NII pour ces derniers, voir les figures 3.3b et c. Si l'on utilise les concepts classiques de sensibilité (proportion de couples expérimentaux effectivement prédits comme tels) et de spécificité (proportion de couples « artificiels » correctement

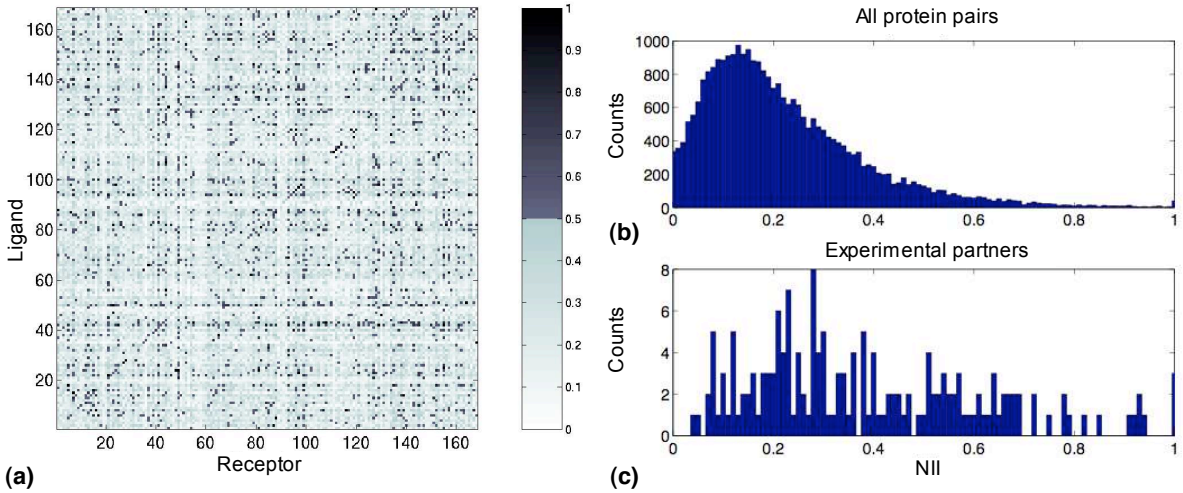


FIG. 3.3 – *(a) Matrice des NII pour le benchmark 2.0, les protéines sont ordonnées de manière à ce que les partenaires expérimentaux soient sur la diagonale. Histogrammes des NII : (b) Pour la totalité des couples protéiques issus du benchmark 2.0, (c) Pour les seuls couples expérimentaux.*

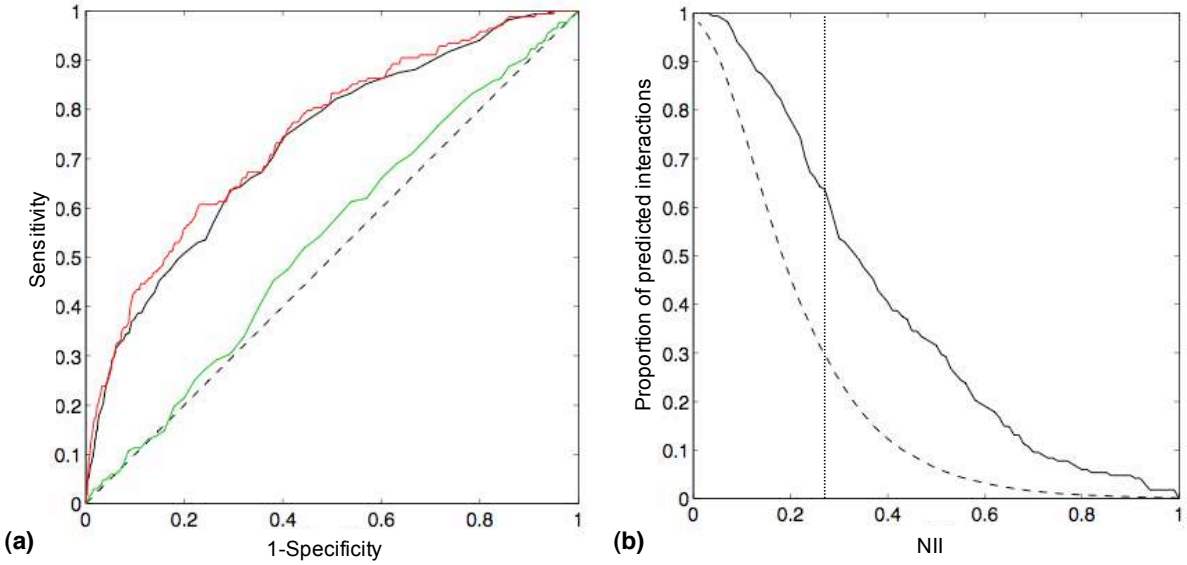


FIG. 3.4 – *(a) Courbe ROC (Receiver Operating Characteristic) pour la prédictions des partenaires d'interaction protéiques à partir du NII (lignes noire et rouge) ou à partir des seules énergies d'interaction (ligne verte). La diagonale indique le comportement d'une prédiction aléatoire. (b) Proportion des interactions protéiques de l'ensemble du benchmark 2.0 prédites, ligne en tirets, et proportion des interactions expérimentales prédites, ligne pleine, en fonction du NII. La verticale en pointillés indique la valeur optimale* $NII = 0.27$.

identifiés), on peut alors évaluer l'efficacité de de notre méthode pour la prédiction des partenaires d'interaction protéique via la valeur de l'aire sous la courbe (AUC) $Sens. = f(1 - Spec.)$ (dite courbe ROC). Selon la définition du NII choisie, L'AUC peut alors atteindre une valeur de 0.75, ce qui est donc nettement supérieur à sa valeur dans le cas de prédictions aléatoires (0.50), ou pour des prédictions basées sur les seules énergies d'interactions (0.52), voir la figure 3.4a. L'optimum des prédictions correspond au point de la courbe ROC le plus éloigné de la diagonale. Cela correspond ici à un NII de 0.27, ce qui entraîne un taux de couverture de 30% des interactions et une sensibilité de 64%, voir la figure3.4b. À titre de comparaison, la méthode de prédiction des partenaires protéiques développée par Yoshikawa *et al.*[106], et qui est basée sur la complémentarité de forme des protéines (dans leur structure liée) aboutit, au mieux, à une AUC de 0.59

On obtient des performances similaires si l'on analyse séparément les trois catégories de complexes, enzyme/inhibiteur (EI), antigène/anticorps (AgAb) et autres (others, O) référencées dans le benchmark 2.0. Cependant si l'on se restreint aux interactions entre protéines appartenant à des catégories « complémentaires » (E+I, Ag+Ab) on observe alors une augmentation du NII moyen (de 0.22 à 0.25) qui signale une meilleur reconnaissance des partenaires.

Le benchmark 2.0 comprend également des complexes de type antigène/anticorps lié (Ag-BAb), pour lesquels le second partenaire est dans sa forme liée (la structure de la forme isolée n'étant pas disponible). Comme on pouvait s'y attendre cette catégorie présente des performances qui sont toujours supérieures à celle du groupe AgAb, ce qui souligne l'importance de la prise en compte des changement structuraux pour la modélisation correcte des interactions protéiques [107].

3.3 Prédiction des sites d'interaction protéique

La limitation majeure de l'exploitation des expériences de docking croisé pour la prédiction des partenaires d'interaction protéique concerne l'utilisation d'informations expérimentales au sujet des résidus présents au niveau de l'interface protéique. Le développement d'approches permettant la prédiction de ces résidus indépendamment de toute donnée expérimentale concernant les partenaires d'interaction potentiels représente donc un point essentiel de ce projet. Celui-ci a été abordé de deux manières différentes, et qui seront par la suite exploitées conjointement. Soit par une approche structurale qui exploite les données issues des calculs de docking croisé, soit par une approche bioinformatique basée sur l'analyse de la conservation des séquences au sein d'une famille protéique.

3.3.1 Apport des calculs de docking croisé

L'analyse des structures dockées produites lors de nos calculs peut apporter une solution partielle au problème de la prédiction des résidus d'interface. En effet, Fernandez-Recio *et al.* [108] ont déjà montré que le docking de protéines n'appartenant pas à un même complexe permet cependant d'identifier un certain nombre des résidus d'interface expérimentaux. Nous avons donc repris le concept de Normalized Interface Propensity (NIP), afin d'évaluer la probabilité pour un résidu de surface de se trouver au niveau d'une interface protéine-proteine. Si l'on considère l'intégralité des interfaces protéiques produites lors du docking systématique d'une protéine P_1 avec l'ensemble des éléments du benchmark 2.0, le NIP d'un résidu i appartenant à P_1 est défini comme la proportion de structures dockées où i est présent au niveau de l'interface. Cette valeur est pondérée par l'énergie des interfaces considérées, afin de favoriser les résidus appartenant à des interfaces de basse énergie, et normée de manière à ce que pour l'ensemble de la protéine $\langle NIP \rangle = 0$ et qu'un $NIP > 0$ signale un résidu susceptible d'appartenir à l'interface, voir [104] pour plus de détails sur le calcul des NIP.

La figure 3.5a montre la prédiction de résidus d'interface en fonction du NIP. Si l'on choisit un cutoff $NIP = 0.0$ (valeur optimale), on aboutit alors à la sélection de 34% des résidus de

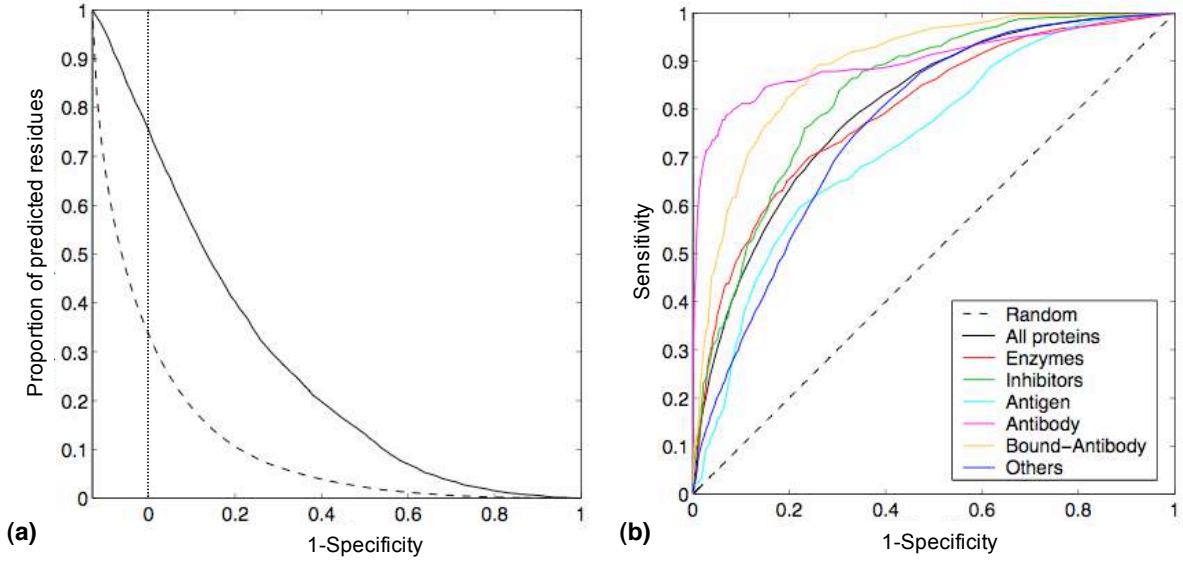


FIG. 3.5 – *(a) Proportion des résidus de surface de l'ensemble du benchmark 2.0 prédits comme des résidus d'interface, ligne en tirets, et proportion des résidues d'interface expérimentaux prédits comme tels, ligne pleine, en fonction du NIP. La verticale en pointillés indique la valeur* $NIP = 0.0$ *(b) Courbe ROC pour la prédictions des résidus d'interface à partir du NIP pour les différentes catégories de protéines du benchmark2.0. La diagonale indique le comportement d'une prédiction aléatoire.*

surface et 75% des résidus d'interface expérimentaux. Comme on peut le voir sur la figure 3.5b, la qualité des prédictions varie alors en fonction du type de protéine considéré. On obtient notamment de meilleurs résultats pour les anticorps (dans leur configuration liée ou non liée) par rapport à l'ensemble de la base, alors que prédictions pour les antigènes sont au contraire dégradées. Ces résultats sont à mettre en rapport avec les données de Kowalsman et Eisentein [109], qui ont mis en évidence la spécificité des interfaces antigéniques, qui sont en général plus difficiles à détecter que celles des anticorps.

3.3.2 Approche phylogénétique

Parallèlement à ces travaux, j'ai travaillé avec l'équipe d'A. Carbone à la mise au point d'un programme de détection des sites fonctionnels dans les protéines, appelé Joint Evolutionary Trees (JET)[110], exploitant la méthode « Evolutionary Trace »[111]. À partir d'un ensemble d'arbres phylogénétiques construits pour une famille de protéines donnée, cette approche permet d'extraire des résidus « trace » de la séquence en acides aminés, ces résidus correspondant à des positions conservées dans les différentes branches des arbres. Des études préalables ont montré que les résidus trace ainsi obtenus forment des clusters dans la structure tridimensionnelle de la

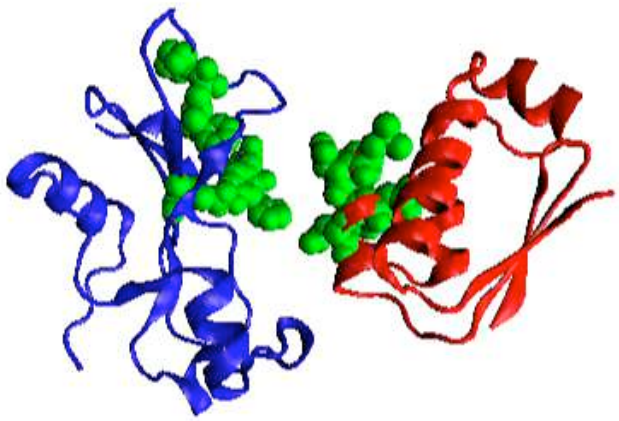


FIG. 3.6 – *Complexe protéique barnase (en bleu)/barstar(en rouge), les résidus d'interface détectés par JET sont représentés en vert.*

protéine et sont localisés au niveau des sites fonctionnels ou des interfaces macromoléculaires, voir la figure 3.6. Cette méthode va donc nous apporter des informations supplémentaires concernant les sites d'interaction à partir de la seule structure primaire d'une protéine et sans aucune donnée sur ses partenaires potentiels.

Contrairement à la méthode ET d'origine qui passe par la construction d'un unique arbre phylogénétique comprenant toutes les séquences homologues obtenues, JET utilise plusieurs arbres, construits chacun à partir d'un sous-ensemble de séquences homologues, et obtient donc des valeurs moyennées pour l'extraction des résidus trace. Cette méthode permet alors d'amplifier les signaux de certains sites d'interaction autrement trop faibles pour être détectés par ET et de prendre en compte des protéines de basse homologie (avec moins de 30% d'identité par rapport à la séquence de référence) lors de la construction des arbres phylogénétiques afin d'améliorer la robustesse des prédictions du programme.

3.4 Conclusions

Aprés avoir été testées séparément, les deux approches (bioinformatique et modélisation) seront appliquées conjointement aux protéines du benchmark 2.0, notamment en intégrant des données produites par JET au calcul des NII entre protéines. Les informations ainsi obtenues sur les interfaces macromoléculaires seront recoupées pour mettre au point une base de données des sites d'interaction protéiques. Les résultats des calculs de docking croisé effectués à grande échelle seront utilisés pour comparer interactions spécifiques (entre partenaires expérimentaux) et non-spécifiques. Les informations obtenues par JET vont également permettre de réduire le coût des calculs de docking en limitant l'exploration des surfaces protéiques aux seuls sites

d'interaction détectés prélablement (ce qui représente une réduction des points de départ nécessaires par un facteur cent). Le gain réalisé en matière de temps de calculs va alors rendre possible l'analyse de bases de données protéiques nettement plus large (comprenant plusieurs milliers de structures).

Les calculs réalisés sur la WCG ont été faits sous les noms de projets Help Cure Muscular Distrophy 1 et 2 (cette seconde partie étant encore en cours jusqu'au printemps 2012). En effet, notre objectif à terme est d'appliquer cette approche pluridisciplinaire à un ensemble de protéines connues pour leur implication dans les maladies neuromusculaires afin d'identifier des partenaires d'association (et inhibiteurs) potentiels au sein des bases de données protéiques.

Chapitre 4

Perspectives

Les travaux présentés dans ce rapport posent les bases de mes recherches à venir. Loin d'être des programmes figés, ProPHet et MAXDo ont vocation à évoluer au cours du temps afin de s'adapter à l'étude de nouveaux systèmes et de venir compléter les informations obtenues par d'autres approches théoriques (modélisations tout atome à l'échelle classique ou quantique...) ou expérimentales (spectroscopie d'absorption, RMN, titration par calorimétrie isotherme...). Les paragraphes suivant décrivent brièvement trois projets ANR (financés ou soumis) dans lesquels je serai impliquée dans les années à venir et qui feront intervenir la modélisation des protéines à l'échelle gros-grain.

4.1 Mécanique des protéines

4.1.1 Mécanisme d'ouverture d'un canal ionique

Les canaux ioniques activés par un ligand externe forment une classe majeure de récepteurs de neurotransmetteurs dans le système nerveux central humain et représentent une cible thérapeutique importante. Les équipes de M. Delarue et P-J. Corringer à l'Institut Pasteur ont récemment résolu par cristallographie la structure tridimensionnelle d'un analogue bactérien du récepteur nicotinique de l'acetylcholine, GLIC, montrant ainsi une transition conformationnelle d'une forme ouverte vers une forme fermée qui serait contrôlée par le pH. Une collaboration avec le LBT a été alors mise en place dans le cadre du **projet ANR Nicochimera** afin de mieux comprendre, en couplant approches expérimentales et théoriques, le mécanisme de régulation de cette transition.

Dans un premier temps, M. Baaden au LBT a effectué des simulations longue durée de dynamique moléculaire classique, qui ont mis en évidence la stabilité du récepteur dans sa structure ouverte à pH 4.6[112]. Les simulations en gros-grain réalisées à l'aide de ProPHet sur

cette structure vont permettre d'établir le paysage mécanique de cette protéine, afin de localiser d'une part des résidus clés pour la transition conformationnelle au cœur du canal ionique, et identifier d'autre part les modes normaux impliqués dans cette transition.

4.1.2 Biocatalyseurs d'oxydation de l'hydrogène pour les piles à combustible

Les hydrogénases Ni-Fe sont des enzymes clés pour la conversion de l'hydrogène en protons. L'objectif du **projet ANR Biopac** est d'exploiter celles-ci en remplacement des catalyseurs chimiques pour les procédés de type pile à combustion. Ce projet implique l'immobilisation des protéines sur diverses électrodes tout en conservant leur activité enzymatiques dans des conditions de température et d'oxydation variées. Dans cette perspective, ProPHet sera modifié pour permettre de modéliser l'influence du contact enzyme/surface sur les propriétés mécaniques et la dynamique de cette dernière. Les informations ainsi obtenues seront ensuite exploitées pour optimiser le mode de fixation des hydrogénases sur les électrode de manière à préserver au mieux leur activité catalytique. Ce projet va également bénéficier des résultats de nos travaux portant sur les globines. En effet, l'application du protocole couplé dynamique moléculaire/ProPHet décrit plus haut va nous permettre d'identifier les résidus clés contrôlant la migration du ligand H_2 au sein de l'enzyme, de mieux comprendre les origines de la résistance à O_2 dans certaines espèces spécifiques et comment préserver cette résistance lors de l'adsorption des protéines sur une électrode.

4.2 Interactions protéiques

Vers une cartographie à l'échelle du génome

Le **projet MAPPING** (Making Accurate Predictions of Protein-protein INteractions on the Genomic scale) se place dans la continuité du projet Decrypthon décrit au chapitre précédent en lui ajoutant une composante expérimentale. Il implique en effet l'équipe de F. Penin, de l'Institut de Biologie et Chimie des Protéines à Lyon, qui va réaliser des expériences de titration par calorimétrie isotherme afin d'obtenir pour la première fois des données thermodynamiques, du type constante d'affinité, pour des protéines ne formant a priori pas de complexe connu expérimentalement. Ces données pourront être comparées aux informations issues des calculs de docking croisé effectués avec MAXDo. Qui plus est, la procédure de docking dans MAXDo sera affinée en intégrant d'une part les données évolutives fournies par JET et ses versions ultérieures développées dans le groupe de Génomique Analytique du Pr. A. Carbone, et d'autre

part en exploitant le modèle gros-grain pour les protéines PALACE développé dans l'équipe de R. Lavery à L'IBCP et qui permet de prendre en compte l'influence de la flexibilité protéique lors de la formation d'un complexe. Notre objectif à terme est la mise au point d'un outil pouvant différencier interactions spécifiques et non spécifiques, et l'installation d'un serveur web d'analyse des interaction protéiques disponible pour l'ensemble de la communauté scientifique.

Collaborations développées dans le cadre des travaux présentés

Équipe du Pr. P. Sebban, Laboratoire de Chimie Physique,
CNRS UMR800, Université Paris-sud
Équipe du Pr. R. Clarke,
Département de Chimie de l'Université de Sydney
Étude de la migration des petits ligands au sein de la neuroglobine
Co-encadrement, pour la partie théorique, de la thèse d'Anthony Bocahut.

Équipe du Pr. A. Carbone, Génomique Analytique,
Génomique des microorganismes, FRE3214, CNRS-Université Pierre et Marie Curie
Prédiction des partenaires et sites d'interaction protéiques par des méthodes de modélisation et bioinformatique
Développement conjoint des logiciels MAXDo et JET.

Pr. G. Venturoli, Laboratoire de Biochimie et Biophysique, Université de Bologne, Italie
Étude du centre réactionnnel de R. Sphaeroides, relation entre les propriétés mécaniques de la protéine et sa dynamique interne
Réalisation de calculs en parallèles des expériences de spectroscopie d'absorption.

Équipe de Biophysique (D. Bensimon et V. Croquette),
Laboratoire de Physique Statistique, UMR8550, ENS, Paris
Projet ANR FonFlon :
Relation fonction/fluctuations structurales dans les systèmes enzymatiques
Réalisation de calculs sur les transitions structurales de l'enzyme Guanylate Kinase.

Bibliographie

[1] Russel, D., Lasker, K., Phillips, J., Schneidman-Duhovny, D., Velazquez-Muriel, J. A., and Sali, A. (2009) *Curr Opin Cell Biol* **21(1)**, 97–108.

[2] Tozzini, V. (2005) *Curr Opin Struct Biol* **15(2)**, 144–150.

[3] G A Voth, (ed.) (2009) Coarse-Graining of Condensed Phase and Biomolecular Systems, CRC Press Taylor and Francis Group, .

[4] Rader, A. J. (2010) *Curr Opin Pharmacol* **10(6)**, 753–759.

[5] Neri, M., Anselmi, C., Cascella, M., Maritan, A., and Carloni, P. (2005) *Phys Rev Lett* **95(21)**, 218102.

[6] Lyman, E., Ytreberg, F. M., and Zuckerman, D. M. (2006) *Phys Rev Lett* **96(2)**, 028105.

[7] Ayton, G. S., Noid, W. G., and Voth, G. A. (2007) *Curr Opin Struct Biol* **17(2)**, 192–198.

[8] Heath, A. P., Kavraki, L. E., and Clementi, C. (2007) *Proteins* **68(3)**, 646–661.

[9] Noid, W. G., Liu, P., Wang, Y., Chu, J.-W., Ayton, G. S., Izvekov, S., Andersen, H. C., and Voth, G. A. (2008) *J Chem Phys* **128(24)**, 244115.

[10] Noid, W. G., Chu, J.-W., Ayton, G. S., Krishna, V., Izvekov, S., Voth, G. A., Das, A., and Andersen, H. C. (2008) *J Chem Phys* **128(24)**, 244114.

[11] Tozzini, V. (2010) *Acc Chem Res* **43(2)**, 220–230.

[12] Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and deVries, A. H. (2007) *J Phys Chem B* **111(27)**, 7812–7824.

[13] Levitt, M. (1976) *J Mol Biol* **104(1)**, 59–107.

[14] Taketomi, H., Ueda, Y., and Go, N. (1975) *Int J Pept Protein Res* **7(6)**, 445–459.

[15] Baker, D. (2000) *Nature* **405(6782)**, 39–42.

[16] Koga, N. and Takada, S. (2001) *J Mol Biol* **313(1)**, 171–180.

[17] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res* **28(1)**, 235–242.

[18] Miyazawa, S. and Jernigan, R. (1985) *Macromolecules* **18(3)**, 534–552.

[19] Tirion, M. (1996) *Phys Rev Lett* **77(9)**, 1905–1908.

[20] Tama, F. and Sanejouand, Y. H. (2001) *Protein Eng* **14(1)**, 1–6.

[21] Kondrashov, D. A., Van Wynsberghe, A. W., Bannen, R. M., Cui, Q., and Phillips, G. N. J. (2007) *Structure* **15(2)**, 169–177.

[22] Sanejouand, Y.-H. Les modes normaux de basse fréquence des protéines. Université Claude Bernard-Lyon 1. Habilitation à Diriger les Recherche. 2007.

[23] Bahar, I., Atilgan, A. R., and Erman, B. (1997) *Fold Des* **2(3)**, 173–181.

[24] Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001) *Biophys J* **80(1)**, 505–515.

[25] Doruker, P., Jernigan, R. L., and Bahar, I. (2002) *J Comput Chem* **23(1)**, 119–127.

[26] Bahar, I. and Jernigan, R. L. (1999) *Biochemistry* **38(12)**, 3478–3490.

[27] Bahar, I., Erman, B., Jernigan, R. L., Atilgan, A. R., and Covell, D. G. (1999) *J Mol Biol* **285(3)**, 1023–1037.

[28] Yang, L.-W. and Bahar, I. (2005) *Structure* **13(6)**, 893–904.

[29] Hinsen, K. (1998) *Proteins* **33(3)**, 417–429.

[30] Hinsen, K., Thomas, A., and Field, M. J. (1999) *Proteins* **34(3)**, 369–382.

[31] Navizet, I., Cailliez, F., and Lavery, R. (2004) *Biophys J* **87(3)**, 1426–1435.

[32] Tama, F., Wriggers, W., and Brooks, C. L. (2002) *J. Mol. Biol.* **321**, 297–305.

[33] Delarue, M. and Sanejouand, Y.-H. (2002) *J Mol Biol* **320(5)**, 1011–1024.

[34] Zacharias, M. (2003) *Protein Sci* **12(6)**, 1271–1282.

[35] Zacharias, M. (2005) *Proteins* **60(2)**, 252–256.

[36] Bastard, K., Prevost, C., and Zacharias, M. (2006) *Proteins* **62(4)**, 956–969.

[37] May, A. and Zacharias, M. (2007) *Proteins* **69(4)**, 774–780.

[38] Saladin, A., Fiorucci, S., Poulain, P., Prevost, C., and Zacharias, M. (2009) *BMC Struct Biol* **9**, 27.

[39] Fiorucci, S. and Zacharias, M. (2010) *Proteins* **78(15)**, 3131–3139.

[40] Daniel, R. M., Dunn, R. V., Finney, J. L., and Smith, J. C. (2003) *Annu Rev Biophys Biomol Struct* **32**, 69–92.

[41] Yuan, Z., Zhao, J., and Wang, Z.-X. (2003) *Protein Eng* **16(2)**, 109–114.

[42] Ermak, D. and McCammon, J. (1978) *Journal of Chemical Physics* **69(4)**, 1352–1360.

[43] Rotne, J. and Prager, S. (1969) *Journal of Chemical Physics* **50(11)**, 4831–&.

[44] Wade, R. C., Davis, M. E., Luty, B. A., Madura, J. D., and McCammon, J. A. (1993) *Biophys J* **64(1)**, 9–15.

[45] Sacquin-Mora, S. and Lavery, R. (2006) *Biophys J* **90(8)**, 2706–2717.

[46] Dias, J. M., Alves, T., Bonifacio, C., Pereira, A. S., Trincao, J., Bourgeois, D., Moura, I., and Romao, M. J. (2004) *Structure* **12(6)**, 961–973.

[47] Ptitsyn, O. B. (1998) *J Mol Biol* **278(3)**, 655–666.

[48] Ptitsyn, O. B. and Ting, K. L. (1999) *J Mol Biol* **291(3)**, 671–682.

[49] Sacquin-Mora, S., Laforet, E., and Lavery, R. (2007) *Proteins* **67(2)**, 350–359.

[50] Charvin, G., Allemand, J., Strick, T., Bensimon, D., and Croquette, V. (2004) *Contemporary physics* **45(5)**, 383–403.

[51] Koster, D. A., Croquette, V., Dekker, C., Shuman, S., and Dekker, N. H. (2005) *Nature* **434(7033)**, 671–674.

[52] Moffitt, J. R., Chemla, Y. R., Smith, S. B., and Bustamante, C. (2008) *Annu Rev Biochem* **77**, 205–228.

[53] Brockwell, D. J. (2007) *Biochem Soc Trans* **35(Pt 6)**, 1564–1568.

[54] Puchner, E. M. and Gaub, H. E. (2009) *Curr Opin Struct Biol* **19(5)**, 605–614.

[55] Galera-Prat, A., Gomez-Sicilia, A., Oberhauser, A. F., Cieplak, M., and Carrion-Vazquez, M. (2010) *Curr Opin Struct Biol* **20(1)**, 63–69.

[56] Kumar, S. and Li, M. S. (2010) *Physics Reports-Review Section of Physics Letters* **486(1-2)**, 1–74.

[57] Puchner, E. M., Alexandrovich, A., Kho, A. L., Hensen, U., Schafer, L. V., Brandmeier, B., Grater, F., Grubmuller, H., Gaub, H. E., and Gautel, M. (2008) *Proc Natl Acad Sci U S A* **105(36)**, 13385–13390.

[58] Oberhauser, A. F. and Carrion-Vazquez, M. (2008) *J Biol Chem* **283(11)**, 6617–6621.

[59] Dietz, H., Berkemeier, F., Bertz, M., and Rief, M. (2006) *Proc Natl Acad Sci U S A* **103(34)**, 12724–12728.

[60] Sacquin-Mora, S. and Lavery, R. (2009) *Chemphyschem* **10(1)**, 115–118.

[61] Brockwell, D. J., Paci, E., Zinober, R. C., Beddard, G. S., Olmsted, P. D., Smith, D. A., Perham, R. N., and Radford, S. E. (2003) *Nat Struct Biol* **10(9)**, 731–737.

[62] Carrion-Vazquez, M., Li, H., Lu, H., Marszalek, P. E., Oberhauser, A. F., and Fernandez, J. M. (2003) *Nat Struct Biol* **10(9)**, 738–743.

[63] Nome, R. A., Zhao, J. M., Hoff, W. D., and Scherer, N. F. (2007) *Proc Natl Acad Sci U S A* **104(52)**, 20799–20804.

[64] Zocchi, G. (2009) *Annu Rev Biophys* **38**, 75–88.

[65] Tseng, C.-Y., Wang, A., Zocchi, G., Rolih, B., and Levine, A. J. (2009) *Phys Rev E Stat Nonlin Soft Matter Phys* **80(6 Pt 1)**, 061912.

[66] Choi, B., Zocchi, G., Wu, Y., Chan, S., and Jeanne Perry, L. (2005) *Phys Rev Lett* **95(7)**, 078102.

[67] Choi, B. and Zocchi, G. (2007) *Biophys J* **92(5)**, 1651–1658.

[68] Tseng, C.-Y., Wang, A., and Zocchi, G. (2010) *European Physics Letters* **91**, 18005.

[69] Sacquin-Mora, S., Delalande, O., and Baaden, M. (2010) *Biophys J* **99(10)**, 3412–3419.

[70] Sacquin-Mora, S., Sebban, P., Derrien, V., Frick, B., Lavery, R., and Alba-Simionesco, C. (2007) *Biochemistry* **46(51)**, 14960–14968.

[71] Paddock, M. L., Feher, G., and Okamura, M. Y. (2003) *FEBS Lett* **555(1)**, 45–50.

[72] Rutherford, A. W. and Faller, P. (2003) *Philos Trans R Soc Lond B Biol Sci* **358(1429)**, 245–253.

[73] Wraight, C. A. (2004) *Front Biosci* **9**, 309–337.

[74] Paddock, M. L., Rongey, S. H., Feher, G., and Okamura, M. Y. (1989) *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6602–6606.

[75] Takahashi, E. and Wraight, C. (1990) *Biochimica et Biophysica Acta* **1020(1)**, 107–111.

[76] Pokkuluri, P. R., Laible, P. D., Deng, Y.-L., Wong, T. N., Hanson, D. K., and Schiffer, M. (2002) *Biochemistry* **41(19)**, 5998–6007.

[77] Veronesi, G., Giachini, L., Francia, F., Mallardi, A., Palazzo, G., Boscherini, F., and Venturoli, G. (2008) *Biophys J* **95(2)**, 814–822.

[78] Francia, F., Malferrari, M., Sacquin-Mora, S., and Venturoli, G. (2009) *Journal of Physical Chemistry B* **113**, 10389–10398.

[79] Burmester, T., Weich, B., Reinhardt, S., and Hankeln, T. (2000) *Nature* **407(6803)**, 520–523.

[80] Burmester, T. and Hankeln, T. (2009) *J Exp Biol* **212(Pt 10)**, 1423–1428.

[81] Vallone, B., Nienhaus, K., Matthes, A., Brunori, M., and Nienhaus, G. U. (2004) *Proc Natl Acad Sci U S A* **101(50)**, 17351–17356.

[82] Hamdane, D., Kiger, L., Dewilde, S., Green, B. N., Pesce, A., Uzan, J., Burmester, T., Hankeln, T., Bolognesi, M., Moens, L., and Marden, M. C. (2003) *J Biol Chem* **278(51)**, 51713–51721.

[83] Brunori, M. and Vallone, B. (2007) *Cell Mol Life Sci* **64(10)**, 1259–1268.

[84] Bocahut, A., Bernad, S., Sebban, P., and Sacquin-Mora, S. (2009) *J Phys Chem B* **113(50)**, 16257–16267.

[85] Scorciapino, M. A., Robertazzi, A., Casu, M., Ruggerone, P., and Ceccarelli, M. (2009) *J Am Chem Soc* **131(33)**, 11825–11832.

[86] Bocahut, A., Bernad, S., Sebban, P., and Sacquin-Mora, S. (2011) *J. Am. Chem. Soc.* **133**, 8753–8761.

[87] Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958) *Nature* **181**, 662–666.

[88] Perutz, M. F. (1960) *Brookhaven Symp Biol* **13**, 165–183.

[89] Jones, S. and Thornton, J. M. (1996) *Proc Natl Acad Sci U S A* **93(1)**, 13–20.

[90] Alberts, B. (1998) *Cell* **92(3)**, 291–294.

[91] Robinson, C. V., Sali, A., and Baumeister, W. (2007) *Nature* **450(7172)**, 973–982.

[92] Wodak, S. J., Pu, S., Vlasblom, J., and Seraphin, B. (2009) *Mol Cell Proteomics* **8(1)**, 3–18.

[93] Shoemaker, B. A. and Panchenko, A. R. (2007) *PLoS Comput Biol* **3(3)**, e42.

[94] Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002) *Nature* **415(6868)**, 141–147.

[95] Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005) *Nature* **433(7025)**, 531–537.

[96] Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.-P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.-L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007) *Mol Syst Biol* **3**, 89.

[97] Huang, H. and Bader, J. S. (2009) *Bioinformatics* **25(3)**, 372–378.

[98] Kuchaiev, O., Rasajski, M., Higham, D. J., and Przulj, N. (2009) *PLoS Comput Biol* **5(8)**, e1000454.

[99] Shoemaker, B. A. and Panchenko, A. R. (2007) *PLoS Comput Biol* **3(4)**, e43.

[100] Gray, J. J. (2006) *Curr Opin Struct Biol* **16(2)**, 183–193.

[101] Pons, C., Grosdidier, S., Solernou, A., Perez-Cano, L., and Fernandez-Recio, J. (2010) *Proteins* **78(1)**, 95–108.

[102] Fernandez-Recio, J. and Sternberg, M. (2010) *Proteins* **78(15)**, 3065–3066.

[103] Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., and Weng, Z. (2005) *Proteins* **60(2)**, 214–216.

[104] Sacquin-Mora, S., Carbone, A., and Lavery, R. (2008) *J Mol Biol* **382(5)**, 1276–1289.

[105] Sacquin-Mora, S., Ponty, Y., Carbone, A., and Lavery, R. (2011) *soumis*.

[106] Yoshikawa, T., Tsukamoto, K., Hourai, Y., and Fukui, K. (2009) *J Chem Inf Model* **49(3)**, 693–703.

[107] Zacharias, M. (2010) *Curr Opin Struct Biol* **20(2)**, 180–186.

[108] Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2004) *J Mol Biol* **335(3)**, 843–865.

[109] Kowalsman, N. and Eisenstein, M. (2009) *Proteins* **77(2)**, 297–318.

[110] Engelen, S., Trojan, L. A., Sacquin-Mora, S., Lavery, R., and Carbone, A. (2009) *PLoS Comput Biol* **5(1)**, e1000267.

[111] Lichtarge, O. and Sowa, M. E. (2002) *Curr Opin Struct Biol* **12(1)**, 21–27.

[112] Bocquet, N., Nury, H., Baaden, M., Le Poupon, C., Changeux, J.-P., Delarue, M., and Corringer, P.-J. (2009) *Nature* **457(7225)**, 111–114.

Deuxième partie

Dossier administratif

5.1 Curriculum Vitae

Sophie Sacquin
épouse Mora
Née le 19 juillet 1978 à Bourg-en-Bresse (Ain).

Adresse professionnelle
Laboratoire de Biochimie Théorique, CNRS UPR9080
Institut de Biologie Physico-Chimique
13 rue Pierre et Marie Curie, 75005 Paris
Tel : 01 58 41 51 65,
Fax : 01 58 41 50 26
Mail : sacquin@ibpc.fr
URL : http ://www.ibpc.fr/sacquin

Fonction actuelle
Depuis Octobre 2010
CR1 CNRS (section 13) rattachée au Laboratoire de Biochimie Théorique, CNRS UPR9080 (Dir. P. Derreumaux)

Fonctions occupées

Oct. 2006-Sept 2010	CR2 CNRS (section 13) rattachée au Laboratoire de Biochimie Théorique, CNRS UPR9080 (Dir. P. Derreumaux)
Sept. 2005-Sept. 2006	Postdoctorant INSERM financée par le projet DECRYPTHON.
Nov. 2004-Août 2005	ATER à l'Université d'Évry Val d'Essonne (Département de Physique).
Nov. 2003-Oct. 2004	Postdoctorant CNRS au Laboratoire de Biochimie Théorique (Institut de Biologie Physico-Chimique, Paris).
Sept. 2001-Oct. 2003	Allocataire de recherche et moniteur à l'Université Paris XI, Orsay.
2000-2001	Quatrième année de scolarité à l'École Normale Supérieure.

Formation Universitaire

2000-2003 **Doctorat de l'Université Paris XI, Orsay**
Fluides Nanoconfinés dans des Systèmes de Basse Symétrie : Simulations et Théorie
Thèse de chimie physique réalisée en cotutelle, sous la direction d'Alain Fuchs, Laboratoire de Chimie Physique (Université Paris XI, Orsay), et de Martin Schoen, Stranski Lab. für Physikalische und Theoretische Chemie (Technische Universität, Berlin, Allemagne).

2000-2001 Diplôme du magistère Interuniversitaire de Chimie de l'ENS.

1999-2000 Agrégation de Sciences Physiques option Chimie (rang 20ème).

1998-1999 DEA de Physico-Chimie Moléculaire à l'Université Paris XI, Orsay (Mention Bien).

1997-1998 Première année à l'École Normale Supérieure (Paris).
Second semestre : Maitrise de Chimie Physique (Mention Assez-Bien).
Premier semestre : Licence de Chimie (Mention Bien).

Juillet 1997 Admission à l'École Normale Supérieure (Paris), concours D/S - Physique-Chimie.

1995-1997 Classes Préparatoires PCSI puis PC* au Lycée Louis-le-Grand (Paris).

Juin 1995 Baccalauréat S au Lycée Louis-le-Grand, Paris (Mention Bien).

5.2 Activités de recherche

Sept. 2010-

Participation au projet ANR Nicochimera (36 mois) :
Étude du mécanisme d'ouverture d'un canal ionique
Collaboration entre le LBT (M. Baaden)
et les équipes de M. Delarue et P.-J. Corringer à l'Institut Pasteur

Participation au projet ANR Biopac (48 mois) :
Biocatalyseurs d'oxydation de l'hydrogène pour les piles à combustible
Projet coordiné par E. Lojou,
Laboratoire de Bioénergétique et ingénierie des protéines
UPR9036 (Marseille)
et impliquant le LBT, l'Institut de Chimie des Surfaces et Interfaces
(Mulhouse) et l'équipe Matériaux divisés du Laboratoire Chimie Provence
(Marseille)

Sept. 2008-

Une étude théorique et expérimentale du fonctionnement de la neuroglobine
Collaboration avec le Laboratoire de Chimie Physique, UMR8000,
Université Paris-sud (P. Sebban et S. Bernad)
Et le département de Chimie de l'Université de Sydney (Pr. Ron Clarke)

Sept. 2008-Sept. 2009

Etude de la dynamique interne du centre réactionnel de R. Sphaeroides
Collaboration avec le Laboratoire de Biochimie-Biophysique,
Université de Bologne, Italie (G. Venturoli)

Oct. 2006-Oct. 2009

Participation au projet ANR FonFlon (36 mois) :
Relation fonction/fluctuation chez les protéines
Collaboration entre le LBT (M. Baaden, O. Delalande),
et le Laboratoire de Physique Satistique-UMR8550, ENS Paris
(D. Bensimon, J.-F. Allemand, F. Mosconi)

Continuation des axes de recherche
Interactions protéiques
et *Mécanique des protéines*

Sept. 2005-sept. 2006 *Interactions protéine-protéine*
Séjour post-doctoral effectué dans le cadre du programme DECRYPTHON (projet d'une durée initiale de 18 mois).
Collaboration entre le LBT (Richard Lavery),
et le groupe d'Analyse Génomique, UMR7238 (Alessandra Carbone).

Nov. 2003-août 2005 *Mécanique interne des protéines*
Séjour post-doctoral au Laboratoire de Biochimie Théorique
(Institut de Biologie Physico-Chimique, Paris)
sous la direction de Richard Lavery.

Janvier-août 2002 et Nov. 2000-août 2001 Séjours dans le groupe de Chimie Théorique de Martin Schoen à la Technische Universität (Berlin, Allemagne).

Sept. 2000-oct. 2003 *Fluides nanoconfinés dans des systèmes de basse symétrie : Simulations et théorie*
Thèse au Laboratoire de Chimie Physique (Université Paris XI, Orsay)
sous la direction d'Alain Fuchs.

Janvier-juin 1999 *Simulations Monte-Carlo d'un mélange de sphères dures dans un milieu confiné*
Stage de DEA sous la direction de Martin Schoen,
Laboratoire de Physique Théorique
(Bergische Universität Wuppertal, Allemagne).

5.3 Enseignement

Sept. 2011 **Workshop, Coarse Grain Methods for Biomolecular Simulations**
Institut Pasteur de Montevideo, Uruguay
Organisation d'un TP de simulation numérique
Coarse-Grain Models for Protein Mechanics

2011 Licence Sciences du vivant, parcours de biologie informatique(28h)
Travaux dirigés de Bioinformatique (L3) à l'Université Paris 7, Denis Diderot

2004-2005 **Attachée Temporaire d'Enseignement et de Recherche à l'Université d'Évry Val d'Essonne (demi-poste)**

DEUG Sciences de la Vie (60h)
Chargée de travaux dirigés et travaux pratiques en physique.
(mécanique, mécanique des fluides, thermodynamique).
Participation à l'élaboration des sujets d'examens et correction des examens.

DEUG MIAS (16h)
Chargée de travaux pratiques en électromagnétisme.

DEUG Sciences de la Matière (48h)
Chargée de travaux pratiques en physique (mécanique et thermodynamique).

2001-2003 **Monitorat à l'Université Paris XI (Centre Scientifique d'Orsay)**

DEUG Sciences de la Vie (56h/an)
Chargée de travaux dirigés, travaux pratiques et colles de chimie.
(thermodynamique, thermochimie, atomistique et molécules).
Participation à l'élaboration des sujets d'examens et correction des examens.

Préparation à l'agrégation interne de Physique et Chimie (10h/an)
Chargée de travaux dirigés en atomistique.

5.4 Encadrement et diffusion de la culture scientifique

Oct. 2011- Coencadrement de la thèse de Nikita Chopra
(avec Chantal Prevost, LBT, CNRS UPR9080)
Caractéristiques structurales du peptide NFL-TBS.40-63 seul ou en complexe avec la tubuline

Sept.2011- Coencadrement du stage post-doctoral de Francesco Oteri
(avec Marc Baaden , LBT, CNRS UPR9080)
dans le cadre du projet ANR BIOPAC

Sept. 2011- Coencadrement du stage post-doctoral de Samuel Murail
(avec Marc Baaden , LBT, CNRS UPR9080)
dans le cadre du projet ANR Nicochimera

Mai 2011 Coorganisation du XVIIème congrès du GGMM
(Groupe de Graphisme et Modélisation Moléculaire)

Sept. 2008- Coencadrement de la thèse d'Anthony Bocahut
(avec Pierre Sebban, Laboratoire de Chimie Physique,
CNRS UMR8000, Université Paris-sud)
Une étude théorique et expérimentale du fonctionnement de la neuroglobine
(deux articles publiés et un article en préparation)

Nov. 2007-Oct. 2009 Coencadrement du stage post-doctoral d'Olivier Delalande
(avec Marc Baaden , LBT, CNRS UPR9080)
Dans le cadre du projet ANR FonFlon
Relation entre les fluctuations structurales de guanylate kinase et son activité biologique
(deux articles publiés)

Janv.-Fev. 2006 Encadrement du stage de M1 de Émilie Laforêt,
Étude de la flexibilité locale des protéines pour la localisation des résidus catalytiques
Master de Biochimie, Biologie Moléculaire et Cellulaire,
Université de Franche-Comté.
(un article publié)

2005-2006 Encadrement du stage de M2 de Ladislas Trojan,
Détection des sites d'interaction protéiques
par la méthode Evolutionnary Trace
Master de Biomathématiques, Université Paris 6.
(un article publié)

5.5 Liste de publications

Publications dans des revues à comité de lecture

1)*Fluids confined by nanopatterned substrates of low symmetry*
S. Sacquin, M. Schoen, and A. H. Fuchs, Mol. Phys. **100**, 2971-2982 (2002).

2)*Fluid phase transitions at chemically heterogeneous, non planar solid substrates : Surface versus confinement effects*
S. Sacquin, M. Schoen, and A. H. Fuchs, J. Chem. Phys. **118**, 1453-1465 (2003).

3)*Nanoscopic liquid bridges exposed to a torsional strain*
S. Sacquin-Mora, A. H. Fuchs, and M. Schoen, Phys. Rev. E **68**, 066103 (2003).

4)*Torsion-induced phase transitions in fluids confined between chemically decorated substrates*
S. Sacquin-Mora, A. H. Fuchs, and M. Schoen, J. Chem. Phys. **121**, 9077-9086 (2004).

5)*Investigating the local flexibility of functional residues in hemoproteins*
S. Sacquin-Mora and R. Lavery, Biophys. J. **90**, 2706-2717 (2006).

6)*Locating the active sites of enzymes using mechanical properties*
S. Sacquin-Mora, E. Laforet, and R. Lavery, Proteins **67**, 350-359 (2007).

7)*Protein mechanics : a route from structure to function*
R. Lavery and S. Sacquin-Mora, J. Biosciences **32**, 891-898 (2007).

8)*Probing the flexibility of the bacterial reaction center :*
The wild-type protein is more rigid than two site-specific mutants
S. Sacquin-Mora, P. Sebban, V. Derrien, B. Frick, R. Lavery, and C. Alba-Simionesco,
Biochemistry **46**, 14960-14968 (2007).

9)*Identification of protein interaction partners and protein-protein interaction sites*
S. Sacquin-Mora, A. Carbone, and R. Lavery, J. Mol. Biol. **382**, 1276-1289 (2008).

10)*Modelling the mechanical response of proteins to anisotropic deformations*
S. Sacquin-Mora and R. Lavery, ChemPhysChem **10**, 115-118 (2009).

11)*Joint Evolutionary Trees :*
A large scale method to predict protein interfaces based on sequence sampling
S. Engelen, L. A. Trojan, S. Sacquin-Mora, R. Lavery and A. Carbone,
PLoS Computational Biology **5**, e1000267 (2009).

12)*Charge recombination kinetics and protein dynamics in wild type*
and carotenoid-less bacterial reaction centers : Studies in trehalose glass
F. Francia, M. Malferrari, S. Sacquin-Mora and G. Venturoli,
J. Phys. Chem. B **113**, 10389-10398 (2009).

13)*Relating the diffusion of small ligands in human neuroglobin*
to its structural and mechanical properties
A. Bocahut, S. Bernad, P. Sebban and S. Sacquin-Mora,
J. Phys. Chem. B **113**, 16257-16267 (2009).

14)*Functional modes and flexibility control the anisotropic response*
of Guanylate Kinase to mechanical stress
S. Sacquin-Mora, O. Delalande and M. Baaden, Biophys. J. **99**, 3412-3419 (2010).

15)*Frontier residues lining internal cavities in globins present specific mechanical properties*
A. Bocahut, S. Bernad, P. Sebban and S. Sacquin-Mora,
J. Am. Chem. Soc. **133**, 9753-8761 (2011).

16)*Enzyme closure and nucleotide binding structurally lock guanylate kinase*
O. Delalande, S. Sacquin-Mora and M. Baaden, Biophys. J. 101, 1440-1449 (2011).

17) *High-Throughput investigation of protein-protein interactions via cross-docking simulations*
S. Sacquin-Mora, Y. Ponty, A. Carbone and R. Lavery (2011), soumis.

18) *Closed loops in protein folding*
S. V. Chintapalli, C. J.R. Illingworth, K. E. Parkes, C. R. Snell, G. J. G. Upton,
S. Sacquin-Mora, Lavery, P.J. Reeves and C. A. Reynolds (2011), soumis.

Chapitres d'ouvrages

1) *Coarse-graining protein mechanics.* R. Lavery and S. Sacquin-Mora,
chapitre de **Coarse-Graining of Condensed Phase and Biomolecular Systems**,
G. Voth (ed.), Taylor and Francis, 317-328 (2009).

Valorisation de la recherche

Novembre 2006 S. Sacquin-Mora, R. Lavery,
Etablissement d'un dossier de valorisation pour le logiciel MAXDo
(Molecular Association via Cross-Docking)
et dépôt à l'Agence de Protection des Programmes.

Communications à des congrès

8th European Biophysics Congress
Budapest, Hongrie, août 2011, poster :
Functional modes and residue flexibility control the anisotropic response.
of Guanylate Kinase to mechanical stress.

Colloque Biologie et santé de l'ANR
Lyon, juillet 2011, poster :
Relation fonction/fluctuation chez les protéines

Molecular Perspectives on Protein-Protein Interactions
San Feliu, Espagne, novembre 2010, poster :
Identification of protein interaction partners and protein-protein interaction sites.

Celebrating Computational Biology
Oxford, Royaume-Uni, septembre 2010, poster :
Functional modes and residue flexibility control the anisotropic response.
of Guanylate Kinase to mechanical stress.

Ist Aegean International Conference on Molecular Recognition
Heraklion, Crète, juin 2010, présentation orale :
Identification of protein interaction partners and protein-protein interaction sites.

Journées Modélisation de Paris
Paris, juin 2010, présentation orale :
Mécanique des protéines soumises à une contrainte externe.

VIII European symposium of the Protein Society
Zürich, Suisse, juin 2009, poster :
Modelling the mechanical response of proteins to anisotropic deformations.

Pushing the boundaries of biomolecular simulation
Ascona, Suisse, juin 2008, poster :
Modelling the mechanical response of proteins to anisotropic deformations.

Grand Challenges in Computational Biology
Barcelone, Espagne, juin 2008, poster :
Identification of protein interaction partners and interaction sites.

Soft, Complex, and Biological Mater Conference
Terrasini, Italie, juillet 2007, poster :
Probing macromolecular mechanics : Heterogeneity and Function.

Multi-Protein Complexes Involved in Cell Regulation
Cambridge, Royaume-Uni, août 2006, poster :
Coupling of cross-docking simulations and Evolutionary Trace methods for the prediction of protein-protein interactions.

Journées Modélisations de l'ENS-ENSCP
Paris, 6 et 7 juin 2006, présentation orale :
Flexibilité des résidus fonctionnels dans les protéines à hème.

M2Cell, Workshop on Modeling from Macromolecules to Cells
Abbaye de Fontevraud, décembre 2005, poster :
Large scale investigation of protein-protein interactions via cross docking simulations and Evolutionary Trace methods.

Thermodynamics 2003
Cambridge, Royaume-Uni, avril 2003, présentation orale :
Fluid phase transitions at chemically heterogeneous, nonplanar, solid substrates : Surface versus confinement effects.

Doktorandenkolloquium des Sfb 448
Schwerin, Allemagne, juin 2002, présentation orale :
Phase behavior of a fluid confined by nanopatterned substrates of low symmetry.

Sixth Liblice Conference on the Statistical Mechanics of Liquids
Spindleruv Mlyn, République Tchèque, juin 2002, poster :
Phase behavior of a fluid confined by nanopatterned substrates of low symmetry.

5.6 Résumés des publications scientifiques

Les pages suivantes reproduisent les résumés de mes publications scientifiques concernant la modélisation des protéines. Ces résumés sont rédigés en langue anglaise.

1)*Investigating the local flexibility of functional residues in hemoproteins*
S. Sacquin-Mora and R. Lavery, Biophys. J. **90**, 2706-2717 (2006).

It is now widely accepted that protein function depends not only on structure, but also on flexibility. However, the way mechanical properties contribute to catalytic mechanisms remains unclear. Here, we propose a method for investigating local flexibility within protein structures that combines a reduced protein representation with Brownian dynamics simulations. An analysis of residue fluctuations during the dynamics simulation yields a rigidity profile for the protein made up of force constants describing the ease of displacing each residue with respect to the rest of the structure. This approach has been applied to the analysis of a set of hemoproteins, one of the functionally most diverse protein families. Six proteins containing one or two heme groups have been studied, paying particular attention to the mechanical properties of the active-site residues. The calculated rigidity profiles show that active site residues are generally associated with high force constants and thus rigidly held in place. This observation also holds for diheme proteins if their mechanical properties are analyzed domain by domain. We note, however, that residues other than those in the active site can also have high force constants, as in the case of residues belonging to the folding nucleus of c-type hemoproteins.

2)*Locating the active sites of enzymes using mechanical properties*
S. Sacquin-Mora, E. Laforet, and R. Lavery, Proteins **67**, 350-359 (2007).

We have applied the calculation of mechanical properties to a dataset of almost 100 enzymes to determine the extent to which catalytic residues have distinct properties. Specifically, we have calculated force constants describing the ease of moving any given amino acid residue with respect to the other residues in the protein. The results show that catalytic residues are invariably associated with high force constants. Choosing an appropriate cutoff enables the detection of roughly 80% of catalytic residues with only 25% of false positives. It is shown that neither multidomain structures, nor the presence or absence of bound ligands hinder successful

detections. It is however noted that active sites near the protein surface are more difficult to detect and that non-catalytic, but structurally key residues may also exhibit high force constants.

3)*Protein mechanics : a route from structure to function*
R. Lavery and S. Sacquin-Mora, J. Biosciences **32**, 891-898 (2007).

In order to better understand the mechanical properties of proteins, we have developed simulation tools which enable these properties to be analysed on a residue-by-residue basis. Although these calculations are relatively expensive with all-atom protein models, good results can be obtained much faster using coarse-grained approaches. The results show that proteins are surprisingly heterogeneous from a mechanical point of view and that functionally important residues often exhibit unusual mechanical behaviour. This finding offers a novel means for detecting functional sites and also potentially provides a route for understanding the links between structure and function in more general terms.

4)*Probing the flexibility of the bacterial reaction center :*
The wild-type protein is more rigid than two site-specific mutants
S. Sacquin-Mora, P. Sebban, V. Derrien, B. Frick, R. Lavery, and C. Alba-Simionesco, Biochemistry **46**, 14960-14968 (2007).

Experimental and theoretical studies have stressed the importance of flexibility for protein function. However, more local studies of protein dynamics, using temperature factors from crystallographic data or elastic models of protein mechanics, suggest that active sites are among the most rigid parts of proteins. We have used quasielastic neutron scattering to study the native reaction center protein from the purple bacterium Rhodobacter sphaeroides, over a temperature range of 4-260 K, in parallel with two nonfunctional mutants both carrying the mutations L212Glu/L213Asp –> Ala/Ala (one mutant carrying, in addition, the M249Ala –> Tyr mutation). The so-called dynamical transition temperature, Td, remains the same for the three proteins around 230 K. Below Td the mean square displacement, u2, and the dynamical structure factor, S(Q,omega), as measured respectively by backscattering and time-of-flight techniques are identical. However, we report that above Td, where anharmonicity and diffusive motions take place, the native protein

is more rigid than the two nonfunctional mutants. The higher flexibility of both mutant proteins is demonstrated by either their higher u2 values or the notable quasielastic broadening of S(Q,omega) that reveals the diffusive nature of the motions involved. Remarkably, we demonstrate here that in proteins, point genetic mutations may notably affect the overall protein dynamics, and this effect can be quantified by neutron scattering. Our results suggest a new direction of investigation for further understanding of the relationship between fast dynamics and activity in proteins. Brownian dynamics simulations we have carried out are consistent with the neutron experiments, suggesting that a rigid core within the native protein is specifically softened by distant point mutations. L212Glu, which is systematically conserved in all photosynthetic bacteria, seems to be one of the key residues that exerts a distant control over the rigidity of the core of the protein.

5)*Identification of protein interaction partners and protein-protein interaction sites*
S. Sacquin-Mora, A. Carbone, and R. Lavery, J. Mol. Biol. **382**, 1276-1289 (2008).

Rigid-body docking has become quite successful in predicting the correct conformations of binary protein complexes, at least when the constituent proteins do not undergo large conformational changes upon binding. However, determining whether two given proteins interact is a more difficult problem. Successful docking procedures often give equally good scores for proteins that do not interact experimentally. This is the case for the multiple minimization approach we use here. An analysis of the results where all proteins within a set are docked with all other proteins (complete cross-docking) shows that the predictions can be greatly improved if the location of the correct binding interface on each protein is known, since the experimental complexes are much more likely to bring these two interfaces into contact, at the same time as yielding good interaction energy scores. While various methods exist for identifying binding interfaces, it is shown that simply studying the interaction of all potential protein pairs within a data set can itself help to identify the correct interfaces.

6)*Modelling the mechanical response of proteins to anisotropic deformations*
S. Sacquin-Mora and R. Lavery, ChemPhysChem **10**, 115-118 (2009).

Using a method for investigating local flexibility on the residue scale within pro-

tein structures, which combines a reduced protein representation with Brownian dynamics simulations, we performed calculations on the green fluorescent protein in order to scan its response to anisotropic deformation. The directional spring constants that are computed from our simulations show a strong agreement with the results obtained via single-molecule experiments. Further calculations underline the importance of the structural elements located between the extension points and of their orientation relative to the pulling direction on the protein stiffness.

7)*Joint Evolutionary Trees :*
A large scale method to predict protein interfaces based on sequence sampling
S. Engelen, L. A. Trojan, S. Sacquin-Mora, R. Lavery and A. Carbone,
PLoS Computational Biology **5**, e1000267 (2009).

The Joint Evolutionary Trees (JET) method detects protein interfaces, the core residues involved in the folding process, and residues susceptible to site-directed mutagenesis and relevant to molecular recognition. The approach, based on the Evolutionary Trace (ET) method, introduces a novel way to treat evolutionary information. Families of homologous sequences are analyzed through a Gibbs-like sampling of distance trees to reduce effects of erroneous multiple alignment and impacts of weakly homologous sequences on distance tree construction. The sampling method makes sequence analysis more sensitive to functional and structural importance of individual residues by avoiding effects of the overrepresentation of highly homologous sequences and improves computational efficiency. A carefully designed clustering method is parametrized on the target structure to detect and extend patches on protein surfaces into predicted interaction sites. Clustering takes into account residues' physical-chemical properties as well as conservation. Large-scale application of JET requires the system to be adjustable for different datasets and to guarantee predictions even if the signal is low. Flexibility was achieved by a careful treatment of the number of retrieved sequences, the amino acid distance between sequences, and the selective thresholds for cluster identification. An iterative version of JET (iJET) that guarantees finding the most likely interface residues is proposed as the appropriate tool for large-scale predictions. Tests are carried out on the Huang database of 62 heterodimer, homodimer, and transient complexes and on 265 interfaces belonging to signal transduction proteins, enzymes, inhibitors, antibodies, antigens, and others. A specific set of proteins chosen for their special functional and structural properties illustrate JET behavior on a large variety of interactions cove-

ring proteins, ligands, DNA, and RNA. JET is compared at a large scale to ET and to Consurf, Rate4Site, siteFiNDER|3D, and SCORECONS on specific structures. A significant improvement in performance and computational efficiency is shown.

8) *Charge recombination kinetics and protein dynamics in wild type and carotenoid-less bacterial reaction centers : Studies in trehalose glass*
F. Francia, M. Malferrari, S. Sacquin-Mora and G. Venturoli,
J. Phys. Chem. B **113**, 10389-10398 (2009).

The coupling between electron transfer and protein dynamics has been investigated in reaction centers (RCs) from the wild type (wt) and the carotenoid-less strain R26 of the photosynthetic bacterium Rhodobacter sphaeroides. Recombination kinetics between the primary photoreduced quinone acceptor (QA-) and photoxidized donor (P+) have been analyzed at room temperature in RCs incorporated into glassy trehalose matrices of different water/sugar ratios. As previously found in R26 RCs, also in the wt RC, upon matrix dehydration, P+QA- recombination accelerates and becomes broadly distributed, reflecting the inhibition of protein relaxation from the dark-adapted to the light-adapted conformation and the hindrance of interconversion between conformational substates. While in wet trehalose matrices (down to approximately one water per trehalose molecule) P+QA- recombination kinetics are essentially coincident in wt and R26 RCs, more extensive dehydration leads to two-times faster and more distributed kinetics in the carotenoid-containing RC, indicating a stronger inhibition of the internal protein dynamics in the wt RC. Coarse-grained Brownian dynamics simulations performed on the two RC structures reveal a markedly larger flexibility of the R26 RC, showing that a rigid core of residues, close to the quinone acceptors, is specifically softened in the absence of the carotenoid. These experimental and computational results concur to indicate that removal of the carotenoid molecule has long-range effects on protein dynamics and that the structural/dynamical coupling between the protein and the glassy matrix depends strongly upon the local mechanical properties of the protein interior. The data also suggest that the conformational change stabilizing P+QA- is localized around the QA binding pocket.

9) *Relating the diffusion of small ligands in human neuroglobin to its structural and mechanical properties*
A. Bocahut, S. Bernad, P. Sebban and S. Sacquin-Mora,

J. Phys. Chem. B **113**, 16257-16267 (2009).

Neuroglobin (Ngb), a recently discovered member of the globin family, is overexpressed in the brain tissues over oxygen deprivation. Unlike more classical globins, such as myoglobin and hemoglobin, it is characterized by a hexacoordinated heme, and its physiological role is still unknown, despite the numerous investigations made on the protein in recent years. Another important specific feature of human Ngb is the presence of two cysteine residues (Cys46 and Cys55), which are known to form an intramolecular disulfide bridge. Since previous work on human Ngb reported that its ligand binding properties could be controlled by the coordination state of the Fe(2+) atom (in the heme moiety) and the redox state of the thiol groups, we choose to develop a simulation approach combining coarse-grain Brownian dynamics and all-atom molecular dynamics and metadynamics. We have studied the diffusion of small ligands (CO, NO, and O(2)) in the globin internal cavity network for various states of human Ngb. Our results show how the structural and mechanical properties of the protein can be related to the ligand migration pathway, which can be extensively modified when changing the thiol's redox state and the iron's coordination state. We suggest that ligand binding is favored in the pentacoordinated species bearing an internal disulfide bridge.

10) *Functional modes and flexibility control the anisotropic response of Guanylate Kinase to mechanical stress*
S. Sacquin-Mora, O. Delalande and M. Baaden, Biophys. J. **99**, 3412-3419 (2010).

The coupling between the mechanical properties of enzymes and their biological activity is a well-established feature that has been the object of numerous experimental and theoretical works. In particular, recent experiments show that enzymatic function can be modulated anisotropically by mechanical stress. We study such phenomena using a method for investigating local flexibility on the residue scale that combines a reduced protein representation with Brownian dynamics simulations. We performed calculations on the enzyme guanylate kinase to study its mechanical response when submitted to anisotropic deformations. The resulting modifications of the protein's rigidity profile can be related to the changes in substrate binding affinity observed experimentally. Further analysis of the principal components of motion of the trajectories shows how the application of a mechanical constraint on the protein can disrupt its dynamics, thus leading to a decrease of the enzyme's

catalytic rate. Eventually, a systematic probe of the protein surface led to the prediction of potential hotspots where the application of an external constraint would produce a large functional response both from the mechanical and dynamical points of view. Such enzyme-engineering approaches open the possibility to tune catalytic function by varying selected external forces.

11) *Frontier residues lining internal cavities in globins present specific mechanical properties*
A. Bocahut, S. Bernad, P. Sebban and S. Sacquin-Mora,
J. Am. Chem. Soc. **133**, 9753-8761 (2011).

The internal cavity matrix of globins plays a key role in their biological function. Previous studies have already highlighted the plasticity of this inner network, which can fluctuate with the proteins breathing motion, and the importance of a few key residues for the regulation of ligand diffusion within the protein. In this Article, we combine all-atom molecular dynamics and coarse-grain Brownian dynamics to establish a complete mechanical landscape for six different globins chain (myoglobin, neuroglobin, cytoglobin, truncated hemoglobin, and chains Îś and Îš of hemoglobin). We show that the rigidity profiles of these proteins can fluctuate along time, and how a limited set of residues present specific mechanical properties that are related to their position at the frontier between internal cavities. Eventually, we postulate the existence of conserved positions within the globin fold, which form a mechanical nucleus located at the center of the cavity network, and whose constituent residues are essential for controlling ligand migration in globins.

12) *Enzyme closure and nucleotide binding structurally lock guanylate kinase*
O. Delalande, S. Sacquin-Mora and M. Baaden, Biophys. J. **101**, 1440-1449 (2011).

We investigate the conformational dynamics and mechanical properties of guanylate kinase (GK) using a multi-scale approach combining high-resolution atomistic molecular dynamics and low resolution Brownian dynamics simulations. The GK enzyme is subject to large conformational changes, leading from an open to a closed form, which are further influenced by the presence of nucleotides. As suggested by recent work on simple coarse-grained models of apo-GK, we primarily focus on GK's closure mechanism with the aim to establish a detailed picture of the hierarchy and chronology of structural events essential for the enzymatic reaction. We have inves-

tigated open vs. closed, apo vs. holo and substrate vs. product-loaded forms of the GK enzyme. Bound ligands significantly modulate the mechanical and dynamical properties of GK and rigidity profiles of open and closed states hint at functionally important differences. Our data emphasizes the role of magnesium, highlights a water channel permitting active site hydration and reveals a structural lock that stabilizes the closed form of the enzyme.

13) *Coarse-graining protein mechanics.* R. Lavery and S. Sacquin-Mora,
chapitre de **Coarse-Graining of Condensed Phase and Biomolecular Systems**,
G. Voth (ed.), Taylor and Francis, 317-328 (2009).

In order to better understand the mechanical properties of proteins, we have developed a technique which enables these properties to be analyzed on a residue-by-residue basis. Although these calculations are relatively expensive with all-atom protein models, good results can be obtained much faster using coarse-grained approaches. One-point per amino acid representations capture the overall mechanical behavior of proteins, but a multi-point representation, taking into account side chain size and orientation, is necessary to detecting finer effects. The results obtained by analyzing the fluctuations of a mean-distance function using Brownian dynamic simulations show that proteins are surprisingly mechanically heterogeneous and that functionally important residues often exhibit unusual mechanical behavior, typically being more rigidly fixed within the overall protein structure than others. This finding offers a new way for detecting functional sites and also potentially provides a route for understanding the links between structure and function in more physical terms.

5.7 Publications représentatives

- *Locating the active sites of enzymes using mechanical properties*
  Proteins **67**, 350-359 (2007).
- *Identification of protein interaction partners and protein-protein interaction sites*
  J. Mol. Biol. **382**, 1276-1289 (2008).
- *Functional modes and flexibility control the anisotropic response of Guanylate Kinase to mechanical stress*
  Biophys. J. **99**, 3412-3419 (2010).
- *Frontier residues lining internal cavities in globins present specific mechanical properties*
  J. Am. Chem. Soc. **133**, 9753-8761 (2011).

PROTEINS: Structure, Function, and Bioinformatics 67:350–359 (2007)

Locating the Active Sites of Enzymes Using Mechanical Properties

Sophie Sacquin-Mora, Émilie Laforet, and Richard Lavery*
Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005 Paris, France

ABSTRACT **We have applied the calculation of mechanical properties to a dataset of almost 100 enzymes to determine the extent to which catalytic residues have distinct properties. Specifically, we have calculated force constants describing the ease of moving any given amino acid residue with respect to the other residues in the protein. The results show that catalytic residues are invariably associated with high force constants. Choosing an appropriate cutoff enables the detection of roughly 80% of catalytic residues with only 25% of false positives. It is shown that neither multidomain structures, nor the presence or absence of bound ligands hinder successful detections. It is however noted that active sites near the protein surface are more difficult to detect and that non-catalytic, but structurally key residues may also exhibit high force constants. Proteins 2007;67:350–359.** © 2007 Wiley-Liss, Inc.

Key words: protein structure; flexibility; catalytic residues; elastic network; Brownian dynamics

INTRODUCTION

As the result of numerous structural genomics programs, many structures are being solved for proteins, which have been identified only by the organism to which they belong and by their amino acid sequence. Since many of these proteins have no known homologs, the vital step of determining their biological function has become one of today's significant challenges.[1–4] Identifying function enables the structural information to be fully exploited and, if it could be obtained from the sequence alone, would also enable the most interesting proteins to be selected for structural study.

Identifying function can be approached from several directions. On the genomic level, it is possible to look for gene fusion events or cooccurrences of the gene in question, implying involvement in a single protein complex or in related steps of a metabolic pathway (for review see Vazquez et al.[5]). On the sequence level, high sequence identities generally imply identical functions; however, functions can diverge from 60% identity[6] and this divergence becomes significant below 30%.[7] Comparison with proteins belonging to the same fold family can also be used,[8] although it is clear that the number of functions largely exceeds the number of folds[9] and that catalytic sites with a given function can change during evolution both in terms of the functional groups involved in catalysis and in terms of their location within the protein scaffold.[10]

More reliable predictions generally involve identifying active site residues, which can then be used to extend sequence searches to lower degrees of homology.[11–13] This subproblem can again be approached from several directions. From sequence data alone, it is possible to identify highly conserved residues that are generally associated with interaction surfaces, although they can also be linked to key roles in the folding pathway.[2,14–17] When structures are available, one can use geometrical criteria such as the existence of local structural patterns,[18–20] the identification of surface clefts,[21–26] or the related measure of proximity to the protein centroid.[27,28] Other key characteristics include the nature of the residues. Roughly, 70% of catalytic residues belong to a group of six amino acids (Arg, Asp, Cys, Glu, His, and Lys). Catalytic residues also generally exhibit high polarities, relatively low solvent exposure, and extensive hydrogen bonding.[29]

It has also been noted that catalytic residues are generally more rigid than others, with lower crystallographic[29,30] or theoretically determined[31] *B*-factors (which are proportional to atomic mean square displacements) and are also often destabilizing elements within the protein architecture, notably from an electrostatic point of view.[32,33] Both the latter characteristics suggest that catalytic residues may have particular mechanical properties within the overall protein structure. While *B*-factors are helpful in characterizing such properties, it has been shown that they largely reflect local environments, being closely related to local atomic packing densities,[34] and they may consequently be insensitive to larger scale features. In a recent study, Yang and Bahar[31] showed that better results could be obtained by looking at mean square displacements related to the collective protein movements represented by one or two of the lowest frequency normal modes.

Our interest in the mechanical properties of biological macromolecules[35] recently led us to look for new ways of

*Correspondence to: Richard Lavery, Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005 Paris, France. E-mail: rlavery@ibpc.fr

Received 17 October 2006; Accepted 24 November 2006

Published online 20 February 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21353

TABLE I. Results for Six Representative Enzymes Drawn From the Dataset

Protein	PDB code	Number of amino acids	Known catalytic residues	Predicted catalytic residues	Total ASA of catalytic residues (Å^2)
Phospholipase A2	1BK9	134	48, 52, 99	48, 52, 99	44
Ricin	1BR6	268	80, 81, 121, 123, 177, 180	80, 81, 121, 123, 177, 180	126
HIV-1 protease	1A30	99+99	A25, A30, B25	A25, A30, B25	33
Plasminogen activator	1A51	244	57, 102, 156, 194, 195	57,102, 156, 194, 195	45
Asv integrase	1A5V	54–199	64, 121, 157	64	151
Human rhinovirus 3C protease	1CQQ	180	40, 71, 145, 147	71, 147	139

analyzing the mechanics of proteins.[36,37] As an alternative to studying mean square fluctuations, which are dominated by local environments, or low frequency normal modes, which are less easy to relate to residue-level properties, we have developed a measurement of the ease of displacing a given residue with respect to the overall protein structure. This involves calculating a force constant to quantify the ease with which the mean distance from a given residue to all other residues in the structure can be modified. This force constant can be obtained directly by perturbing the mean distance using constrained energy minimization[36] or deduced from the fluctuations of the mean distance observed during a molecular dynamics trajectory.[36,37] We have shown that both these procedures lead to very similar results and also that an all-atom representation of the protein can be replaced with a coarse-grain elastic network model[38,39] with very little loss of precision.

We recently applied this approach to a set of hemoproteins.[37] In line with the discussion earlier, we noted that the residues surrounding the heme groups and involved in the catalytic mechanism were generally associated with high force constants (and thus exhibited low mobilities). These results encouraged us to apply our approach to a much larger and more varied group of enzymes to see whether this new mechanical probe could become a useful guide to key catalytic residues. The results presented here involve studies of a dataset of almost 100 proteins belonging to different enzymatic classes. The combination of an intermediate level elastic network representation (which models side chain volume and conformation as well as the backbone fold) with rapid Brownian dynamics simulations enables us to obtain force constants for each residue in the protein dataset. It is shown that this approach can successfully identify almost 80% of the catalytic residues with only 25% of false attributions. Again in connection with the discussion earlier, we discuss how active site locations, the presence of multiple structural domains, or the presence or absence of bound ligands can influence the results. Finally, we discuss how the predictions might be improved by pruning structurally-important, but noncatalytic, residues out of the high force constant group.

MATERIALS AND METHODS

Dataset

Calculations have been carried out on the broad data set of almost 100 enzymes created by Yang and Bahar.[31] This data covers two groups of enzymes, representative of the six EC classes: 93 nonhomologous monomeric enzymes and five multimeric enzymes. These enzymes are listed in the online supplemental data of Ref. 31. Catalytic residues within the set are defined using the criteria established by Bartlett et al.,[29] that is, a residue is catalytic if: (i) it is directly involved in a catalytic function; (ii) it affects residues or water molecules directly involved in catalysis; (iii) it can stabilize a transient intermediate; (iv) it interacts with a substrate or cofactor that facilitates the local chemical reaction. Yang and Bahar divided the 98 enzymes studied into two sets: 24 involving inhibitor binding (and also including the five multimeric enzymes), which allows both catalytic and ligand binding residues to be defined (see Table I of Ref. 31) and 74 monomeric enzymes whose catalytic residues are listed in the Catalytic Site Atlas.[40]

Calculating Force Constants for Each Residue

The local flexibility of the dataset of enzymes was studied using a technique we have previously applied to hemoproteins.[37] This involves determining how difficult it is to move a given residue with respect to the other residues in the protein, or, more precisely, calculating a force constant for changing the mean distance from the probed residue to all other residues. This information can be obtained by constrained variation of the mean distance using energy minimization or by observing the fluctuations of the same mean distance during dynamic simulations of the protein. Earlier studies have shown that very similar results are obtained whichever method is employed and that the results are also largely independent of the protein representation, whether atomic or coarse-grained.[36,37]

Here, we use the technique we have previously applied to the study of hemoproteins, which has the advantage of allowing force constants to be obtained rapidly (typically, a few hours of computation per protein on a standard PC

workstation). The technique involves Brownian dynamics simulations on a coarse-grained protein representation, related to that used in Gaussian network models.[41,42] In contrast to the most common coarse-grain models, this representation, developed by Zacharias,[43] involves 2–3 pseudoatoms for each residue. Each amino acid has one pseudoatom at the Cα position. Small side chains (excepting glycine) have a second pseudoatom at the geometric center of the heavy atoms of the side chain, while larger side chains (Arg, Gln, Glu, His, Lys, Met, Trp, and Tyr) have a pseudoatom at the center of the Cβ-Cγ bond and a third pseudoatom at the geometrical center of the heavy atoms of the side chain atoms beyond Cγ. This somewhat more fine-grained representation has already proved useful in protein–protein docking studies,[44,45] since it models the space occupied by each residue more accurately and takes account of the varying side chain volumes and conformations of the various amino acid residues.

Interactions between the pseudoatoms of the Zacharias representation are treated using the elastic network model, that is, points falling below a cutoff distance of 9 Å are joined with harmonic springs. All springs have the same force constant and are assumed to be relaxed in the reference conformation of the protein. The spring force constant is taken here to be 0.6 kcal mol^{-1} $Å^{-2}$, a value somewhat smaller than in one point-per-residue coarse-grained models (usually set to roughly 1.0 kcal mol^{-1} $Å^{-2}$ [31,39,46]), in order to offset the higher spring density of this representation. The pseudoatom positions for each enzyme studied were derived from crystallographic coordinates contained in the protein data bank.[47] Following our earlier study,[37] prosthetic groups were not included in these representations. This choice enables the intrinsic mechanical properties of each protein to be analyzed independently of the nature and position of any bound ligands. Our work on hemoproteins showed that ligands as large as the heme group actually have little influence on the calculated force constants. Whether this remains true for other ligands encountered in the present data set will be discussed later.

Mechanical properties were obtained from 50,000 steps of Brownian dynamics simulations at 300 K, which were carried out for each protein in the dataset. The simulations were analyzed in terms of the fluctuations of the mean distance between each pseudoatom belonging to a given amino acid residue and the pseudoatoms belonging to the remaining residues of the protein. The inverse of these fluctuations yields an effective force constant k_i describing the ease of moving a given pseudoatom with respect to the overall protein structure.

$$k_i = \frac{3k_BT}{\left\langle \left(d_i - \langle d_i \rangle\right)^2 \right\rangle}$$

where $\langle\rangle$ denotes an average taken over the whole simulation and $d_i = <d_{ij}>_{j^*}$ is the average distance from particle i to the other particles j in the protein (the sum over j^* implies the exclusion of pseudoatoms belonging to residue i). The distances between the Cα pseudoatom of residue i and the Cα pseudoatoms of the adjacent residues $i-1$ and $i+1$ are also excluded since the corresponding distances are virtually constant. The force constant for each residue k is simply the average of the force constants for all its constituent pseudoatoms i. We will use the term "rigidity profile" to describe the ordered set of force constants for all the residues of a given protein.

Analysis of the Results

To determine the links between catalytic activity and mechanical properties, we have tested the ability of our calculated force constants to predict the experimentally determined results for the proteins composing the present dataset. However, since the average magnitude of the calculated force constants varies as a function of the size of the protein studied, smaller proteins generally exhibiting smaller force constants,[36,37] we begin by normalizing the force constants k by converting them to Z-scores, k', which are given in units of variance $\sigma(k)$ with respect to the mean, $\langle k \rangle$, of the distribution of values for each protein:

$$k' = \frac{k - \langle k \rangle}{\sigma(k)}$$

Consequently, higher than average force constants (for each protein) yield positive Z-scores and lower than average values yield negative Z-scores.

Next, we will assume that there is an ideal cutoff value k'_c for the normalized force constants, which will separate them into two groups, PP (predicted positive) with force constants above the cutoff, predicted to be catalytically active and PN (predicted negative) with force constants below the cutoff, predicted to be noncatalytic. To evaluate the quality of our predictions, we use the classical notions of sensitivity and specificity. Sensitivity (Sen.) is defined as the number of correctly predicted catalytic residues (true positives, TP) divided by the total number of experimentally defined catalytic residues (T). Specificity (Spe.) is defined as the number of correctly predicted noncatalytic residues (true negatives, TN) divided by the total number of experimentally defined noncatalytic residues (F). Optimal predictions would have both sensitivity and specificity equal to unity. If this cannot be achieved, the best result is obtained by minimizing the error function:

$$\text{Err.} = \sqrt{(1 - \text{Sen.})^2 + (1 - \text{Spe.})^2}$$

Lastly, a selection of enzyme active sites was analyzed in terms of the accessible surface area of their catalytic residues. These calculations were performed with the ACCESS program[48] using its default parameters.

RESULTS

As we saw in our earlier study of hemoproteins, the force constant profiles of proteins generally give more sharply varying results than the corresponding B-factors. An example of this is shown in Figure 1 for the case of

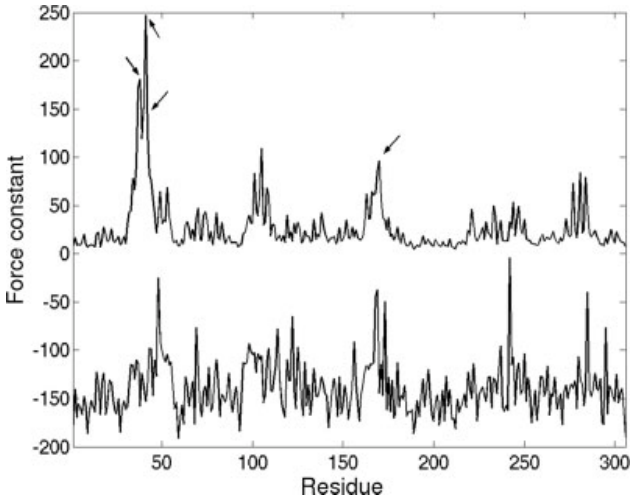


Fig. 1. Rigidity profiles for horseradish peroxidase (1ATJ). Upper curve: force constants, arrows indicate catalytic or ligand-binding residues. Lower curve: inverse of *B*-factors (fitted using proportionality constant and a vertical offset of -200 kcal mol^{-1} Å^{-2}). The force constants in this figure and in Figures 4, 6(B), and 7 are in kcal mol^{-1} Å^{-2} (Note: 1 kcal mol^{-1} Å^{-2} = 0.07 nN Å^{-1}).

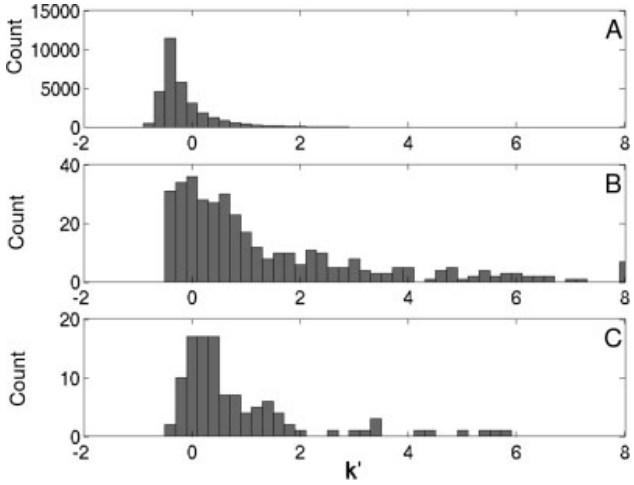
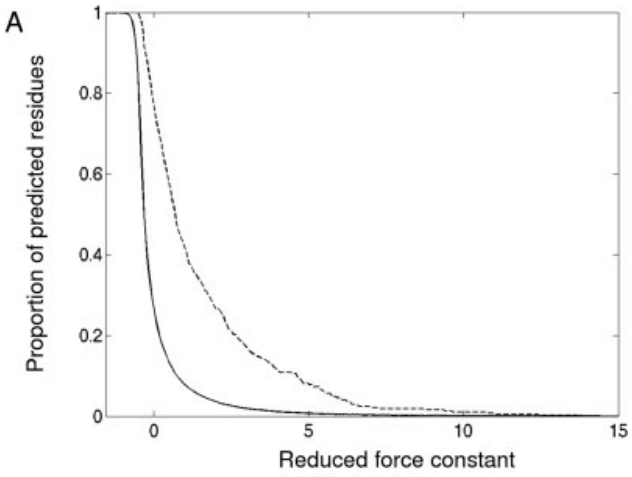


Fig. 2. Distribution of normalized force constants (k') for (**A**) all residues in the data set of 93 monomeric enzymes, (**B**) only the catalytic residues, and (**C**) the ligand-binding residues. The histogram counts were obtained for bin widths of 0.2.



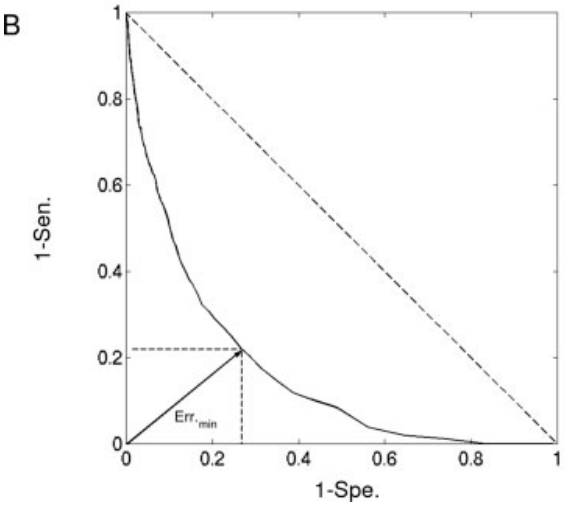


Fig. 3. **A**: Fraction of residues detected as a function of the normalized force constant cutoff k'_c (solid line), contrasted with the fraction of catalytic residues detected with the same criterion (dashed line). **B**: Evolution of the sensitivity and the selectivity as a function of the normalized force constant cutoff.

horseradish peroxidase (pdb code 1ATJ[49]). Our force constants are plotted against the inverse of the *B*-factors, calculated with the same protein model, so that the most positive values in either curve refer to the highest rigidities. This result, which turns out to be valid for the extensive protein database studied here, makes it easier to isolate a small number of potentially catalytic residues (indicated by arrows in the example shown in Fig. 1).

Overall Results for the Enzyme Dataset

Figure 2 illustrates the distribution of the normalized force constants for all the residues in the dataset [Fig. 2(A)] and for the subset of catalytic residues [Fig. 2(B)]. We can see that distribution B has become more strongly skewed with many residues shifted to high force constants, a fact reflected by the mean k', which has increased to 1.5 (compared to 0.0, by definition, for the complete set of residues). While it is not the main aim of this study, we note, in line with the findings of Yang and Bahar,[31] that the ligand binding residues also show a tendency towards higher values [Fig. 2(C)], although the shift is less marked than for the catalytic residues ($\langle k' \rangle = 0.9$).

For a more quantitative view of the results for the entire dataset, Figure 3(A) shows the proportion of residues selected as a function of the normalized force constant cutoff ($k' > k'_c$) (solid line) and the corresponding proportion of catalytic residues selected (dashed line). Figure 3(B) shows variation of the sensitivity and the specificity of the predictions as a function of the cutoff. The dashed diago-

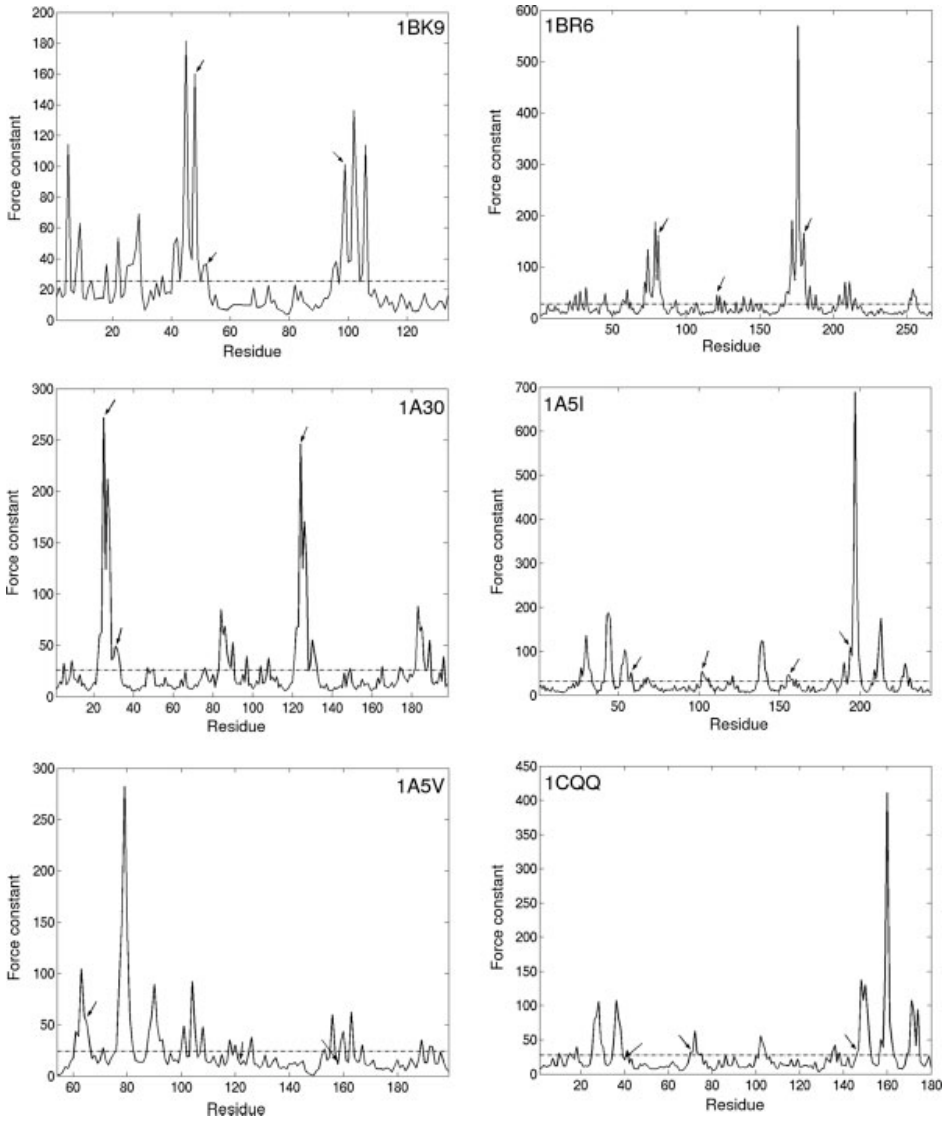


Fig. 4. Rigidity profiles for a set of six sample proteins. The dotted lines indicate the average force constant for each protein (equivalent to the $k_c' = 0$ cutoff) and the arrows indicate the catalytic residues: Phospholipase A2 (1BK9[50]); Ricin (1BR6[51]); HIV-1 Protease (1A30[52]); Plasminogen activator (1A5I[53]); Asv integrase (1A5V[54]); Human rhinovirus 3C protease (1CQQ[55]).

nal corresponds to a random selection of residues. The shortest distance from the sensitivity-selectivity line yields the lowest error estimate for our approach. The optimum cutoff k_c' occurs at a Z-score $= 0$ corresponding to a selection of 28% of the total set of residues. With this cutoff, the error is 0.35, corresponding to a sensitivity of 0.78 and a specificity of 0.74.

Specific Cases

We will illustrate our results using the six specific enzymes listed in Table I. Their rigidity profiles can be found in Figure 4. These enzymes were also studied by Yang and Bahar, which allows us to compare the two sets of results in detail. As can be seen in Figure 4, most of the catalytic residues (indicated by arrows) have force constants considerably above the average. The data in Table I shows that our approach with $k_c' = 0$ is able to detect nearly all of the experimentally identified catalytic residues for these proteins, with a few exceptions, which are discussed later. For the cases where comparison was possible, this measure appears to perform generally better than the mobility criteria of Yang and Bahar.[31]

Surface Active Sites

Among the rigidity profiles presented in Figure 4, it can be seen that the catalytic residues of ASV integrase (1A5V) and rhinovirus protease (1CQQ) generally lie much closer to (or below) the $k_c' = 0$ cutoff than those of

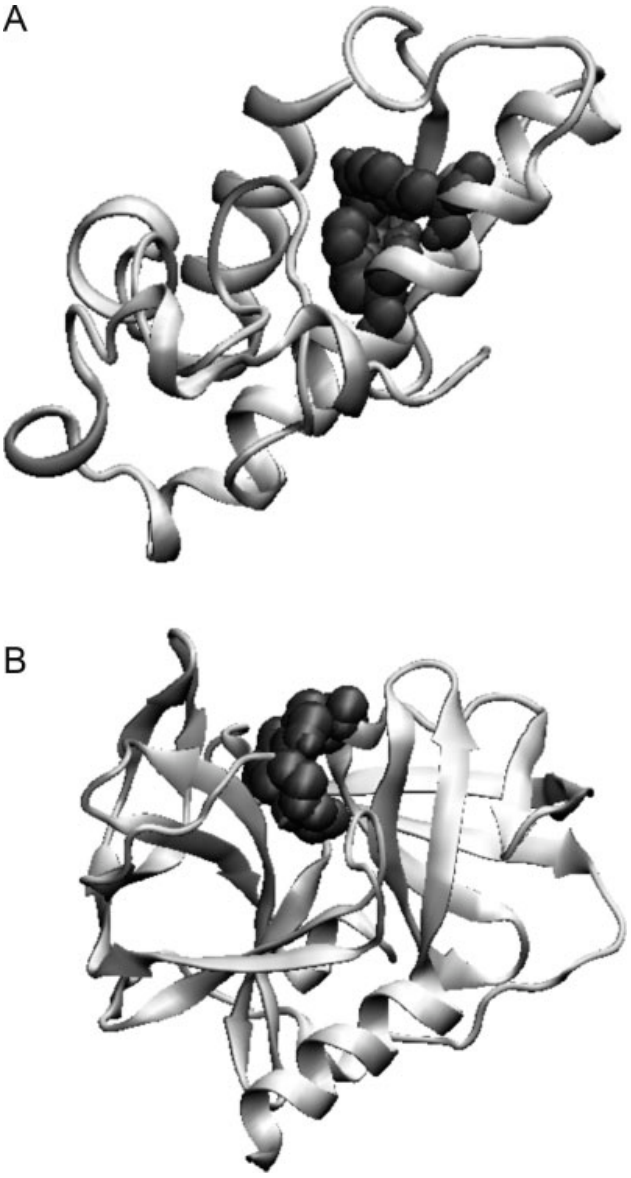


Fig. 5. Simplified representations of phospholipase A2 (1BK9) and human rhinovirus 3C protease (1CQQ) showing the catalytic residues with dark grey van der Waals volumes. The images in Figures 5, 6(A), and 8(A) were prepared using visual molecular dynamics.[56]

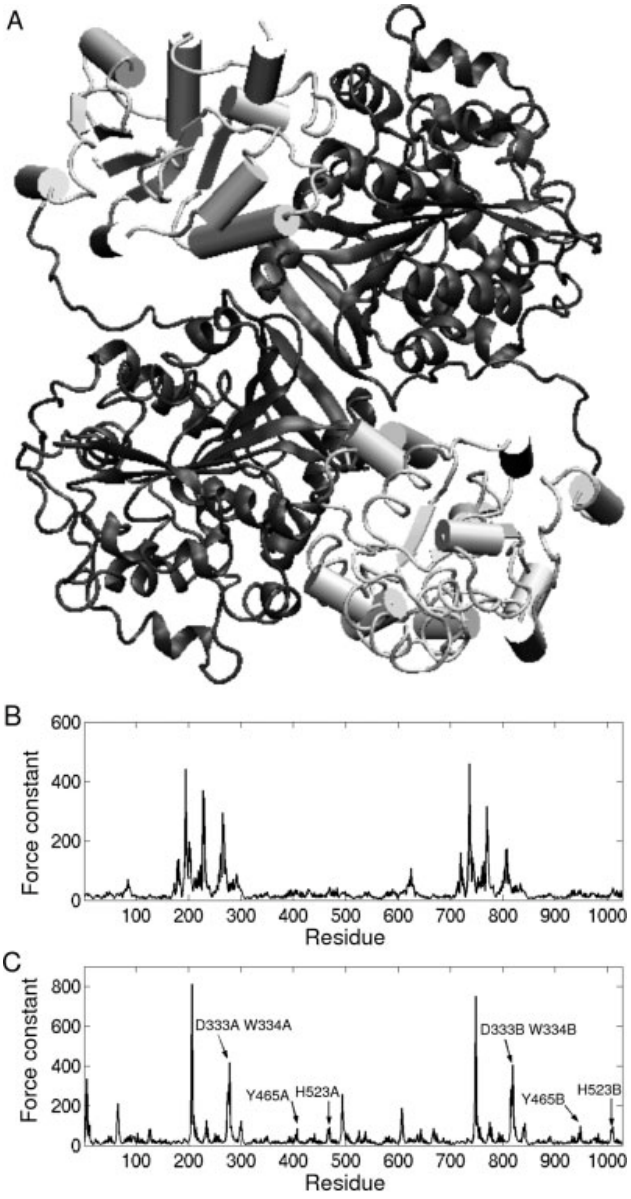


Fig. 6. **A**: Simplified representation of epoxide hydrolase (1CR6) showing the two domains of each chain: light grey for the N-terminal domain and dark grey for the C-terminal domain. **B**: Rigidity profile before domain separation. **C**: Rigidity profile after domain separation, with arrows indicating the new peaks corresponding to catalytic residues.

the other proteins. In these two cases we fail to identify the majority of the active residues, missing two out of three for 1A5V and two out of four for 1CQQ. Interestingly, Yang and Bahar's approach had similar difficulties with these enzymes.[31] The explanation appears to lie in the location of the active sites. If we calculate the accessibilities of the catalytic residues, the results in Table I show that the total values for 1A5V and 1CQQ are the largest for this subset of enzymes. This is due to the fact that the active sites of both these proteins lie on the surface, rather than within a cleft, as illustrated in Figure 5, where rhinovirus protease (1CQQ) is compared with phospholipase A2 (1BK9). In fact, all the catalytic residues of 1A5V and 1CQQ have at least 5% relative accessibilities when compared to the corresponding isolated residues.[57] Surface exposed active sites may therefore lead to lower rigidities than buried sites. For 1A5V and 1CQQ, this seems to be confirmed by the experimental observation that the active sites of both these enzymes undergo considerable structural rearrangement upon substrate binding.[54,55]

Multidomain Proteins

It has been recognized for some time that motions between the subunits of multidomain proteins lead to

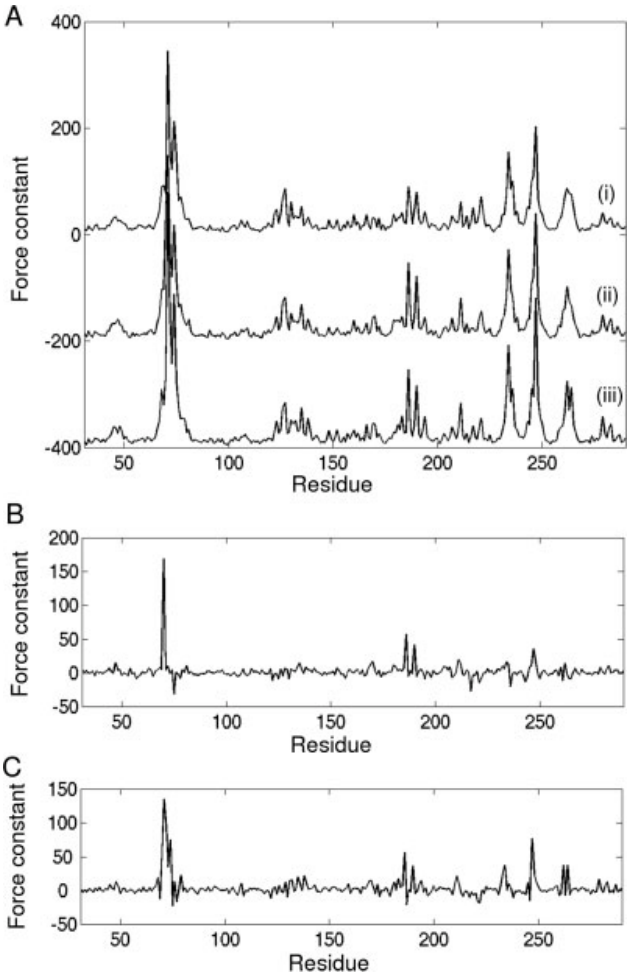


Fig. 7. **A**: Rigidity profiles for β-lactamase: (i) the unliganded conformation (3BLM); (ii) the liganded conformation (1BLC); (iii) the liganded conformation and including the ligand in the elastic network representation. **B**: The difference profile for curves (ii) to (i). **C**: The difference profile for curves (iii)–(i).

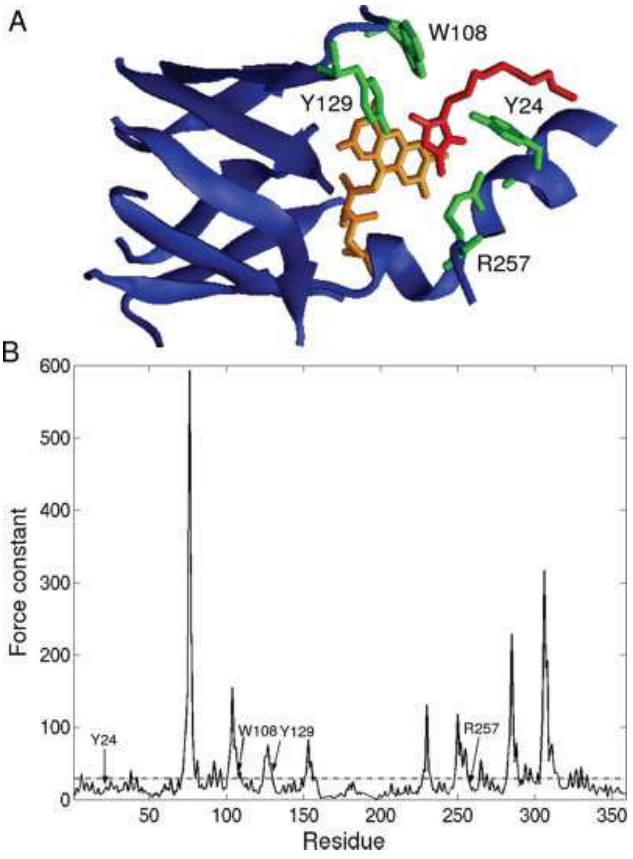


Fig. 8. **A**: Simplified representation of the active site and β-barrel of glycolate oxidase (1AL8). Catalytic residues are shown in green, the flavin mononucleotide group in orange, and the 3-decyl-2,5-dioxo-4-hydroxy-3-pyrroline inhibitor in red. **B**: Rigidity profile of glycolate oxidase, the dotted line indicates the average value of the force constants and the arrows indicate the catalytic residues.

lower *B*-factors for the residues at the hinge points.[58,59] In our approach this is reflected by higher force constants for residues belonging to the domain boundaries.[36] Although it is clear that removing a domain from its overall protein context is likely to radically change its mechanical properties,[60] a simple extension of our approach enables us to study the internal mechanics of a domain without changing its structural context. This extension simply involves modifying the mean distance used to study each residue so that it refers only to other residues within the same domain. This change in no way affects the elastic network representation of the protein in question or the connectivity between multiple domains. We have termed this approach domain separation.[37]

Within the present enzyme dataset, domain separation also turned out to be necessary for proteins built up from nonsymmetric domains and containing more than one active site. The most striking example is epoxide hydrolase (1CR6[61]), a dimeric enzyme where each chain is divided into an N-terminal domain [Arg 4-Gly 218, shown in light grey in Fig. 6(A)] and a catalytic C-terminal domain [Val 235-Ala 544, shown in dark grey in Fig. 6(A)]. Force constants calculated for the protein as a whole [Fig. 6(B)] yield a rigidity profile where all the peaks are located in the two inter-domain regions, and fail to identify any of the catalytic residues. Similar problems were encountered by Yang and Bahar for this enzyme.[31] However, if we apply domain separation (using the domains defined in Ref. 61), we obtain a rigidity profile presenting new peaks above the $k'_c = 0$ cutoff [Fig. 6(C)], which include all 10 of the experimentally identified catalytic residues (with k' values ranging from 0.4 to 5.1).

Effect of Ligand Binding on the Rigidity Profile

As stated in the methodology section, all the force constants presented here correspond to calculations on isolated proteins, even when the structures employed were obtained for proteins in the presence of bound ligands. Our earlier results on hemoproteins showed that even a ligand as large as the heme group had relatively little impact on the calculated rigidity profiles.[37] We thus

concluded that the observed rigidity of the active site residues was an intrinsic property of the protein architecture and not a property induced by the ligand stabilizing an otherwise flexible region.

This conclusion turns out to hold for the enzyme dataset studied here. Although ligands (represented within the elastic network model with a density of pseudoatoms compatible with that used for the amino acids) can marginally change the rigidity profile, they do not change the fact that the catalytic residues already show higher than average force constants in the *apo* protein structure. This is illustrated in Figure 7, which shows the rigidity profiles for the unliganded 3BLM,[62] and liganded 1BLC,[63] forms of β-lactamase and, in the latter case, with or without elastic network nodes for its ligand *N*-(2-oxy-3-hydroxybutyl)-*N*-(3-oxy-transpropenyl)amine. As illustrated by the difference profiles in Figs. 7(B,C), neither the change in conformation corresponding to ligand binding nor the presence of the ligand itself have a significant impact on the selection of high force constant residues (although ligand binding residues can naturally be further rigidified by the presence of the ligand, as in the case of Ser 70 for our β-lactamase example).

Structurally Important Residues

The enzyme dataset studied here shows that residues other than those involved in catalysis can also have higher than average force constants. This situation was already encountered in our studies of hemoproteins, where we showed that highly conserved, structurally important residues constituting the so-called folding nucleus within the cytochrome *c* family[64] were also detected as peaks in our rigidity profiles.[37] An example of this effect within the present dataset is shown in Figure 8. This concerns glycolate oxidase (1AL8), where the active site involves residues within the loops connected to an eight-stranded β-barrel. The barrel fold of this protein appears in the rigidity profile as a characteristic series of eight regularly spaced rigidity peaks, which are associated with considerably higher force constants than the catalytic residues (indicated by arrows). Once again, experimental data attests to the relative flexibility of the catalytic residues in this protein.[65]

These results imply that a comprehensive study of the different fold families should make it possible to detect structurally important residues and, subsequently, to edit these residues out of the rigidity profiles of proteins from each family. We are currently working on achieving this goal.

CONCLUSIONS

The aim of this study was to test whether force constants reflecting the ease of displacing a given amino acid within a given protein structure could be used to detect catalytic sites. The results obtained for a dataset of 98 proteins belonging to a wide variety of enzymatic classes suggest that this mechanical property is indeed useful. Using an appropriately chosen cutoff, we are able to detect 78% of the experimentally identified catalytic residues with only 26% of false positives. These results were achieved using a simple elastic network model and an intermediate coarse-grain protein representation, which allows side chain volume and conformation to be taken into account, while being computationally more manageable than all-atom models.

We have overcome difficulties related to hinge regions in multidomain proteins using a so-called domain separation strategy, which enables the internal mechanics of individual domains to be probed without changing the structural environment constituted by the complete protein. It is also found that ligands bound to the enzyme active sites and structural changes between the *apo* and *holo* forms of enzymes produce only limited changes in the calculated rigidity profiles and have little impact on our predictions. This implies that the rigidity of the active site residues is an intrinsic property of the protein architecture and offers the hope of being able to detect active sites in structures modeled by homology, where the position and the nature of the binding ligands may be unknown.

Problems which remain to be solved involve the observation that active sites close to the surface of the protein, although they are relatively rare, seem to have less marked rigidities than more deeply buried sites and also the fact that residues other than those involved in enzymatic catalysis (such as key residues in folding pathways) can also exhibit high rigidities. A systematic analysis of the mechanical profiles of the different families of protein folds should hopefully enable this category of residues to be identified.

Finally, it is interesting to note that although it has become a paradigm that enzymes must be flexible to carry out their catalysis, different methodological approaches now seem to clearly point to the need for rigidly maintaining the relative position of these residues with respect to the overall protein structure. It is also worth noting that the rigidity of catalytic residues determined from elastic network models appears to be compatible with the observations that these residues are involved in extensive hydrogen bonding networks[29] and could also be related to their energetically destabilizing role,[32] although neither of these features are explicitly represented by the elastic model.

REFERENCES

1. Teichmann SA, Murzin AG, Chothia C. Determination of protein function, evolution and interactions by structural genomics. Curr Opin Struct Biol 2001;11:354–363.
2. Chelliah V, Chen L, Blundell TL, Lovell SC. Distinguishing structural and functional restraints in evolution in order to identify interaction sites. J Mol Biol 2004;342:1487–1504.
3. Pellegrini-Calace M, Soro S, Tramontano A. Revisiting the prediction of protein function at CASP6. FEBS J 2006;273:2977–2983.
4. Soro S, Tramontano A. The prediction of protein function at CASP6. Proteins: Struct Funct Bioinform 2005;61:201–213.
5. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 2003;21:697–700.
6. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 2003;333:863–882.
7. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307:1113–1143.

8. Dietmann S, Holm L. Identification of homology in protein structure classification. Nat Struct Biol 2001;8:953–957.
9. Anantharaman V, Aravind L, Koonin EV. Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. Curr Opin Chem Biol 2003;7:12–20.
10. Todd AE, Orengo CA, Thornton JM. Plasticity of enzyme active sites. Trends Biochem Sci 2002;27:419–426.
11. George RA, Spriggs RV, Bartlett GJ, Gutteridge A, MacArthur MW, Porter CT, Al-Lazikani B, Thornton JM, Swindells MB. Effective function annotation through catalytic residue conservation. Proc Natl Acad Sci USA 2005;102:12299–12304.
12. Panchenko AR, Kondrashov F, Bryant S. Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci 2004;13:884–892.
13. Gutteridge A, Bartlett GJ, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 2003;330:719–734.
14. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001;311:395–408.
15. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
16. Mihalek I, Res I, Lichtarge O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. J Mol Biol 2004;336:1265–1282.
17. Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci 1993;9:745–756.
18. Fischer D, Bachar O, Nussinov R, Wolfson H. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. J Biomol Struct Dyn 1992;9:769–789.
19. Jambon M, Imberty A, Deleage G, Geourjon C. A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 2003;52:137–145.
20. Torrance JW, Bartlett GJ, Porter CT, Thornton JM. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol 2005;347:565–581.
21. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci 1996;5:2438–2452.
22. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 1997;15:359–363, 389.
23. Levitt DG, Banaszak LJ. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 1992;10:229–234.
24. Brady GP, Jr, Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 2000;14:383–401.
25. Peters KP, Fauck J, Frommel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
26. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM. A method for localizing ligand binding pockets in protein structures. Proteins: Struct Funct Bioinform 2006;62:479–488.
27. Ben-Shimon A, Eisenstein M. Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. J Mol Biol 2005;351:309–326.
28. del Sol A, Fujihashi H, Amoros D, Nussinov R. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. Protein Sci 2006;15:2120–2128.
29. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324: 105–121.
30. Yuan Z, Zhao J, Wang ZX. Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng 2003;16:109–114.
31. Yang LW, Bahar I. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure (Camb) 2005;13:893–904.
32. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 2001;312:885–896.
33. Ota M, Kinoshita K, Nishikawa K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J Mol Biol 2003;327:1053–1064.
34. Halle B. Flexibility and packing in proteins. Proc Natl Acad Sci USA 2002;99:1274–1279.
35. Lavery R, Lebrun A, Allemand J-F, Bensimon D, Croquette V. Structure and mechanics of single biomolecules: experiment and simulation. J Phys (Cond Mat) 2002;14:R383–R414.
36. Navizet I, Cailliez F, Lavery R. Probing protein mechanics: residue-level properties and their use in defining domains. Biophys J 2004;87:1426–1435.
37. Sacquin-Mora S, Lavery R. Investigating the local flexibility of functional residues in hemoproteins. Biophys J 2006;90:2706–2717.
38. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 1997;2:173–181.
39. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 2001;80:505–515.
40. Porter CT, Bartlett GJ, Thornton JM. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 2004;32:D129–D133.
41. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 1996;77:1905–1908.
42. Tozzini V. Coarse-grained models for proteins. Curr Opin Struct Biol 2005;15:144–150.
43. Zacharias M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci 2003;12: 1271–1282.
44. Zacharias M. ATTRAACT: protein-protein docking in CAPRI using a reduced protein model. Proteins: Struct Funct Bioinform 2005;60:252–256.
45. Bastard K, Prevost C, Zacharias M. Accounting for loop flexibility during protein-protein docking. Proteins: Struct Funct Bioinforma 2006;62:956–969.
46. Doruker P, Jernigan RL, Bahar I. Dynamics of large proteins through hierarchical levels of coarse-grained structures. J Comput Chem 2002;23:119–127.
47. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The protein data bank. Acta Crystallogr D Biol Crystallogr 2002;58 (Pt 6):899–907.
48. Hubbard SJ. ACCESS: a program for calculating accessibilities. Department of Biochemistry and Molecular Biology, University College of London, 1992.
49. Gajhede M, Schuller DJ, Henriksen A, Smith AT, Poulos TL. Crystal structure of horseradish peroxidase C at 2.15 A resolution. Nat Struct Biol 1997;4:1032–1038.
50. Zhao HY, Tang L, Wang XQ, Zhou YC, Lin ZJ. Structure of a snake venom phospholipase A(2) modified by p-bromo-phenacyl-bromide. Toxicon 1998;36:875–886.
51. Yan XJ, Hollis T, Svinth M, Day P, Monzingo AF, Milne GWA, Robertus JD. Structure-based identification of a ricin inhibitor. J Mol Biol 1997;266:1043–1049.
52. Louis JM, Dyda F, Nashed NT, Kimmel AR, Davies DR. Hydrophilic peptides derived from the transframe region of Gag-Pol inhibit the HIV-1 protease. Biochemistry 1998;37:2105–2110.
53. Renatus M, Stubbs MT, Huber R, Bringmann P, Donner P, Schleuning WD, Bode W. Catalytic domain structure of vampire bat plasminogen activator: a molecular paradigm for proteolysis without activation cleavage. Biochemistry 1997;36:13483–13493.
54. Lubkowski J, Yang F, Alexandratos J, Wlodawer A, Zhao H, Burke TR, Neamati N, Pommier Y, Merkel G, Skalka AM. Structure of the catalytic domain of avian sarcoma virus integrase with a bound HIV-1 integrase-targeted inhibitor. Proc Natl Acad Sci USA 1998;95:4831–4836.
55. Matthews DA, Dragovich PS, Webber SE, Fuhrman SA, Patick AK, Zalman LS, Hendrickson TF, Love RA, Prins TJ, Marakovits JT, Zhou R, Tikhe J, Ford CE, Meador JW, Ferre RA, Brown EL, Binford SL, Brothers MA, DeLisle DM, Worland ST. Structure-

assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes. Proc Natl Acad Sci USA 1999;96:11000–11007.
56. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14:33–38, 27–38.
57. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196:641–656.
58. Bahar I, Jernigan RL. Cooperative fluctuations and subunit communication in tryptophan synthase. Biochemistry 1999;38:3478–3490.
59. Isin B, Doruker P, Bahar I. Functional motions of influenza virus hemagglutinin: a structure-based analytical approach. Biophys J 2002;82:569–581.
60. Kurkcuoglu O, Jernigan RL, Doruker P. Loop motions of triosephosphate isomerase observed with elastic networks. Biochemistry 2006;45:1173–1182.
61. Argiriadi MA, Morisseau C, Hammock BD, Christianson DW. Detoxification of environmental mutagens and carcinogens: structure, mechanism, and evolution of liver epoxide hydrolase. Proc Natl Acad Sci USA 1999;96:10637–10642.
62. Herzberg O. Refined crystal-structure of β-lactamase from *Staphylococcus aureus* Pc1 at 2.0-a resolution. J Mol Biol 1991;217:701–719.
63. Chen CCH, Herzberg O. Inhibition of β-lactamase by clavulanate-trapped intermediates in cryocrystallographic studies. J Mol Biol 1992;224:1103–1113.
64. Ptitsyn OB. Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? J Mol Biol 1998;278:655–666.
65. Stenberg K, Lindqvist Y. Three-dimensional structures of glycolate oxidase with bound active-site inhibitors. Protein Sci 1997;6:1009–1015.

doi:10.1016/j.jmb.2008.08.002 J. Mol. Biol. (2008) 382, 1276–1289

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Identification of Protein Interaction Partners and Protein–Protein Interaction Sites

Sophie Sacquin-Mora[1]*, Alessandra Carbone[2,3] and Richard Lavery[4]

[1]*Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005 Paris, France*

[2]*Département d'Informatique, Université Pierre et Marie Curie - Paris 6, UMR S511, 91 Blv de l'Hopital, 75013 Paris, France*

[3]*Génomique Analytique, INSERM U511, 75013 Paris, France*

[4]*Institut de Biologie et Chimie des Protéines, CNRS UMR 5086 / IFR 128 / Université de Lyon 7 passage du Vercors, 69367 Lyon, France*

Received 19 March 2008; received in revised form 10 July 2008; accepted 1 August 2008
Available online 7 August 2008

Edited by M. Sternberg

Rigid-body docking has become quite successful in predicting the correct conformations of binary protein complexes, at least when the constituent proteins do not undergo large conformational changes upon binding. However, determining whether two given proteins interact is a more difficult problem. Successful docking procedures often give equally good scores for proteins that do not interact experimentally. This is the case for the multiple minimization approach we use here. An analysis of the results where all proteins within a set are docked with all other proteins (complete cross-docking) shows that the predictions can be greatly improved if the location of the correct binding interface on each protein is known, since the experimental complexes are much more likely to bring these two interfaces into contact, at the same time as yielding good interaction energy scores. While various methods exist for identifying binding interfaces, it is shown that simply studying the interaction of all potential protein pairs within a data set can itself help to identify the correct interfaces.

Keywords: protein–protein interactions; docking simulations; coarse-grained model; binding site; binding interface

Introduction

Protein–protein interactions are crucial in most biological processes, including metabolism, signaling, gene expression, and immune responses. Moreover, the availability of complete genome sequences and high-throughput analysis techniques have broadened the focus from a single interaction to the whole proteome, and have made the identification of functional protein complexes, and a better understanding of how they form in the crowded cellular environment, a major goal of biology.[1–3]

Many experimental approaches, including yeast two-hybrid analysis, mass spectroscopy and affinity purification,[4–8] have been developed for the detection and identification of interacting proteins in the cellular context, leading to a wealth of data and the development of protein interaction databases.[9–11] Various bioinformatics approaches have also been used to identify interactions including gene clustering and phylogenetic profiling[12] All these methods, however, have their drawbacks and can result in considerable numbers of false positives and negatives.[13] These methods also identify interactions without providing the structure of the corresponding complex, whereas this structure is often a key element in understanding function.

Molecular modeling offers an alternative to these approaches which, if successful, could help in identifying relevant interactions, at the same time

*Corresponding author. E-mail address: sacquin@ibpc.fr.

Abbreviations used: CC-D, complete cross-docking; PDB, Protein Data Bank.

as providing structural models of the corresponding complexes and clarifying the physical principles behind the complex formation.

While protein–protein docking has become a major goal for biophysics and computational biology over the last 30 years,[14–20] docking algorithms have largely been restricted to determining the conformation of complexes between protein partners that are known to interact. This problem is now being solved more and more successfully, especially when complex formation does not lead to major conformational changes in the interacting partners.[17,21] However, almost no attention has been given to the problem of using docking to identify true interacting partners.

Identifying interacting partners within an arbitrary set of proteins is clearly difficult. Here, we attack this problem via complete cross-docking (CC-D), which involves performing docking calculations on all the possible protein pairs within a given dataset and not only on protein pairs that have already been identified experimentally as forming complexes, and therefore N^2 docking trials for N proteins. To our knowledge, this is the first time that such calculations have been carried out in a systematic way. Previous related studies have been limited to looking at the interaction of a single protein (lysozyme) with three potential receptors (chymotrypsin, cytochrome f and UDG),[22] at the competition of small ligands for a single enzymatic binding site (sometimes termed cross-docking),[23–28] or at the docking of various conformations of the receptor and ligand proteins for a single complex.[29–33]

We have used CC-D with a set of 12 proteins known to form six binary complexes. We used a rigid-body docking algorithm combined with a coarse-grain protein representation to test all potential interactions and showed that while this approach predicts good conformations for the experimentally known partners, it completely fails to identify these partners amongst the numerous alternative interactions when considering the protein interaction energy alone. However, adding accurate data on the location of the correct interface residues greatly improves the scoring function and allows the identification of the experimental partner for each protein in the set. We show also that CC-D can itself provide information on the correct binding interfaces, and can consequently improve predictions.

Results

The MAXDo program, (see Materials and Methods) has been applied to a test set of six binary protein complexes (see Table 1) comprising 12 distinct proteins with sizes ranging from 50 to > 500 residues. Here further references to these proteins use their name or the PDB code of the complex they belong to, plus the chain ID of the protein (Table 1). For example, ***1BRS***-A and ***1BRS***-D refer to barnase (A) and barstar (D) in the barnase–barstar complex ***1BRS***. The coordinates for the bound and unbound conformations of both receptor and ligand proteins, are available in the Protein Data Bank, and they belong to the docking benchmarks that were developed by Chen *et al*.[34 35]

For this preliminary study we have chosen five complexes that correspond to enzyme–inhibitor interactions and one that is an enzyme–activator complex. Each protein was docked on all the proteins of the dataset (including itself). Since the receptor and the ligand proteins have distinct roles in our docking algorithm, every pair of proteins A and B was studied twice, with first A and then B being treated as the receptor. Except for ***1GRN***, all the complexes of our dataset belong to the "rigid-body" category of the docking benchmark.[35] In ***1GRN***, complex formation leads to a 1.22 Å root-mean-square deviation (rmsd) change in the C_α atoms of the interface residues. Although the proteins we study generally undergo only minor backbone conformational changes upon association, there can be many side chain reorientations. We have consequently done two series of tests, interacting either the bound or the unbound conformations, to evaluate the impact of these changes.

Simple docking and energy maps

In order to validate our docking algorithm, we first tested it on the experimentally known complexes. For each complex, the method was able to predict the position of the ligand protein correctly with respect to its receptor with an rmsd of the C_α pseudoatoms

Table 1. Summary of the protein complexes investigated in this study

Complex[a] (bound structures)	PDB 1[a] (unbound structure)	PDB 2[a] (unbound structure)	Protein 1[b]	Protein 2[b]
1BRS(A:D)	1A2P(B)	1A19(A)	Barnase (110)	Barstar (89)
2PTC(E:I)	2PTN	6PTI	β-Trypsin (245)	Pancreatic trypsin inhibitor (PTI) (57)
1FSS(A:B)	2ACE(E)	1FSC	Snake venom acetylcholinesterase (535)	Fasciculin II (61)
2TEC(E:I)	1THM	2TEC(I)	Thermitase (279)	Eglin C (63)
1UGH(E:I)	1AKZ	1UGI(A)	Human Uracil-DNA glycosylase (223)	Inhibitor (UDGI) (84)
1GRN(A:B)	1A4R(A)	1RGP	CDC42 GTPase (200)	CDC42 GAP (199)

[a] PDB[55] code for the crystal structure used in this study with the chain IDs in parenthesis.
[b] Number of residues of the protein in parenthesis.

< 3 Å. One of these optimally docked conformations, for the barnase–barstar complex, is shown in Fig. 1 and the corresponding energy map is the first one shown in Fig. 2. Note that the energy maps are drawn so that the experimental binding site on the receptor protein falls in the middle of the map.

Cross-docking with barnase as receptor

We started our search for experimentally identified protein complexes by comparing the energy maps for barnase as receptor with all 12 potential ligand proteins (Fig. 2). These results show that it is not possible to distinguish the correct barnase–barstar complex on this basis. As seen in Fig. 3, all 12 ligand proteins lead to similar ranges for the interaction energies, although there is a tendency for the optimal interaction energy to strengthen as the size of the ligand increases.[36] However, it can be noted in Fig. 2 that most ligand proteins have a clearly defined energy minimum close to the experimental binding site of barstar on the barnase receptor. The only exception to this rule occurs for acetylcholinesterase (***1FSS-A***), which, due to its size (545 residues), cannot approach barnase very closely.

Complete cross-docking (CC-D) on the full dataset

CC-D data for the full dataset is presented in Fig. 4a in the form of a matrix where each square corresponds to a given ligand–receptor complex (each row defining the ligand and each column defining the receptor). The rows and columns have been ordered so that the squares on the trailing diagonal correspond to the experimental complexes. We have also separated the larger protein of each experimental pair (barnase, β-trypsin, acetylcholinesterase, thermitase, uracil-DNA glycosylase, and CDC42 GTPase) from the smaller (barstar, PTI, fasciculin, eglin C, UDGI, and CDC42 GAP). The large proteins are placed first in each row and last in each column, so that interactions between two large proteins occur in the upper left-hand quadrant of the matrix, and interactions between small proteins occur in the lower right-hand quadrant.

The squares in Fig. 4a show the optimal binding energy of each complex using colors, with blue representing the most stable complexes (lowest interaction energies) and red the least stable. This figure demonstrates strikingly that, as in the case of barnase, it is not possible to determine the correct interacting proteins on the basis of interaction energy. Almost all of the most favorable low-energy squares (colored blue) lie off the trailing diagonal — the only exception being for the DNA glycosylase–inhibitor complex (***1UGH***). We can again see the role of size, since the lowest energies (blue) tend to occur in the upper left-hand quadrant of the matrix and the highest (red) in the lower right-hand quadrant.

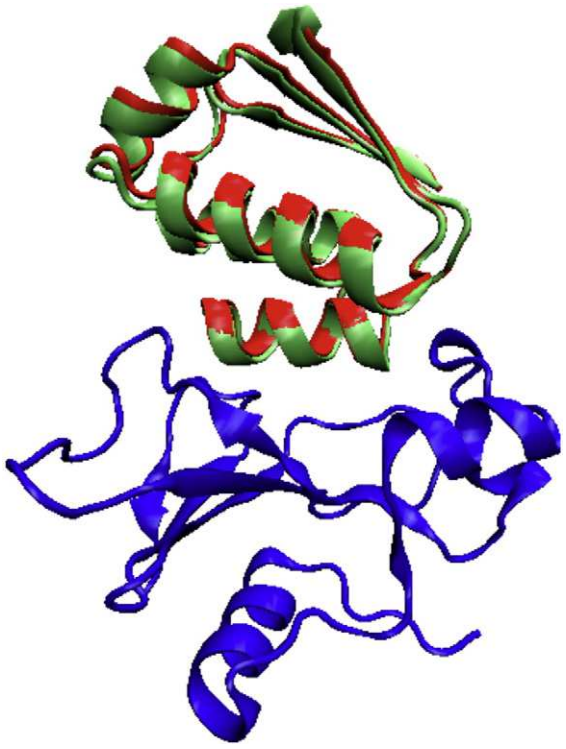


Fig. 1. Comparison of the energy-minimized (green) and experimental (red) positions of barstar in the barnase/barstar complex. Barnase is plotted in blue. The C_α rmsd between the experimental and calculated position is 0.69 Å. The molecular graphics were prepared using VMD. [60]

Buried surface area

One possible explanation for the poor CC-D results could be that our interaction energy does not take account of the desolvation necessary to form a protein complex. This term should be at least approximately proportional to the buried surface area formed when the complex is created. Studies of known complexes have shown that this quantity is typically in the range 1500 – 2700 $Å^2$ (with an average of 1950 $Å^2$).[37] Is it possible that incorrect protein partners lead to much smaller interfaces? In fact, this turns out not to be the case.

Figure 4b shows our cross-docking matrix with squares representing the buried surface areas calculated using the program NACCESS.[38] Once again, it is impossible to separate the experimental complexes (along the trailing diagonal) from the others. The only clear result is that, as might be expected, larger proteins (top left) yield larger interfaces than small proteins (bottom right).

Binding interfaces

Faced with these results, it is necessary to understand what else could be used to characterize a "good" complex. We recall our first cross-docking trial with barnase, in which nearly all ligand proteins had favorable interactions near the experimental binding interface of barnase (even if these were not always global energy minima on the receptor surface). Is it possible that this result is more general?

In order to see to what extent a given complex involves the experimental binding interfaces of the two partners, we started by defining the interface residues as those losing at least 10% of their solvent-accessible surface area upon forming the experimental complex.[22] For a given complex (which now generally does not involve experimentally interacting partners), we determine what fraction of the interface is composed of residues belonging to the experimentally identified interface residues (abbreviated as *FIR*) for the receptor protein (FIR_{rec}) and for the ligand protein (FIR_{lig}). The overall fraction for the complex is defined as:

$$FIR = FIR_{rec} \times FIR_{lig}$$

For every protein pair P_1-P_2, we then calculate an energy-weighted, optimal interface score as:

$$II_{P_1P_2} = \max(FIR)_{P_1P_2} \times E_{tot}(\max(FIR))$$

where max(*FIR*) is the highest value of *FIR* obtained for all the calculated conformations of the complex and E_{tot} the interaction energy of the corresponding conformation. Since all the resulting interaction energies had negative values, we could define a normalized value as:

$$NII_{P_1P_2} = \frac{II^2_{P_1P_2}}{\min(II_{P_1P_j})_{Pj\in P} \times \min(II_{P_jP_2})_{Pj\in P}}$$

where P_n are the 12 proteins of our dataset. *NII* varies between 0 and 1. Values close to zero imply that the two proteins in question cannot form an interface involving a significant fraction of the experimentally identified interface residues or that the "best" interface is associated with a poor interaction energy. Values close to 1 imply interfaces formed from correct residues with good interaction energies.

If we now recalculate our cross-docking matrix using these values, it is immediately clear that the experimental complexes can be detected easily via the blue squares along the trailing diagonal in Fig. 5a. As could be expected from the results described above, all experimental complexes lead to *NII* = 1. All other complexes have lower, often much lower, values.

It is worth noting that both elements specifying *NII* (i.e., correct interface residues and good interaction

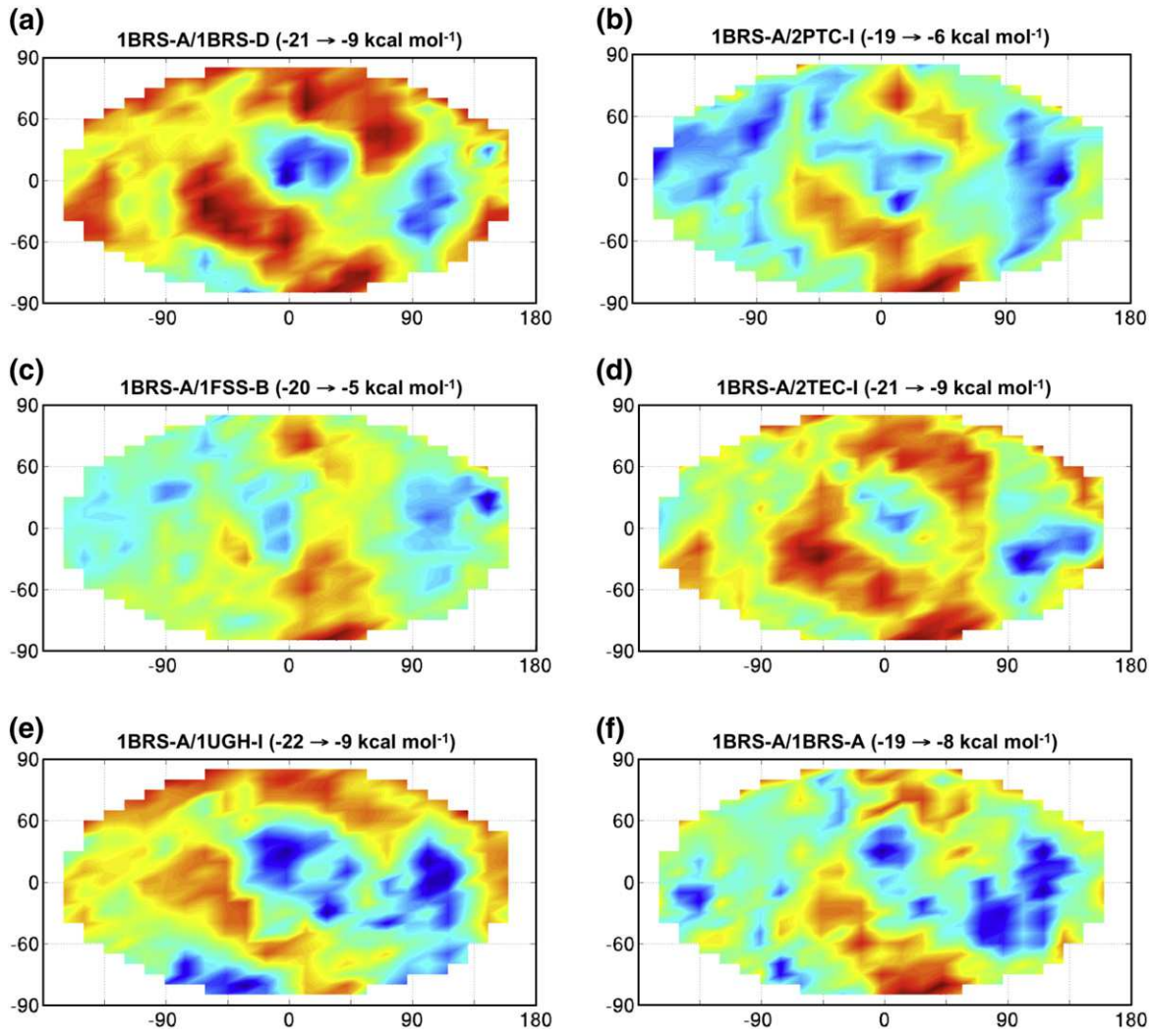


Fig. 2 (*legend on next page*)

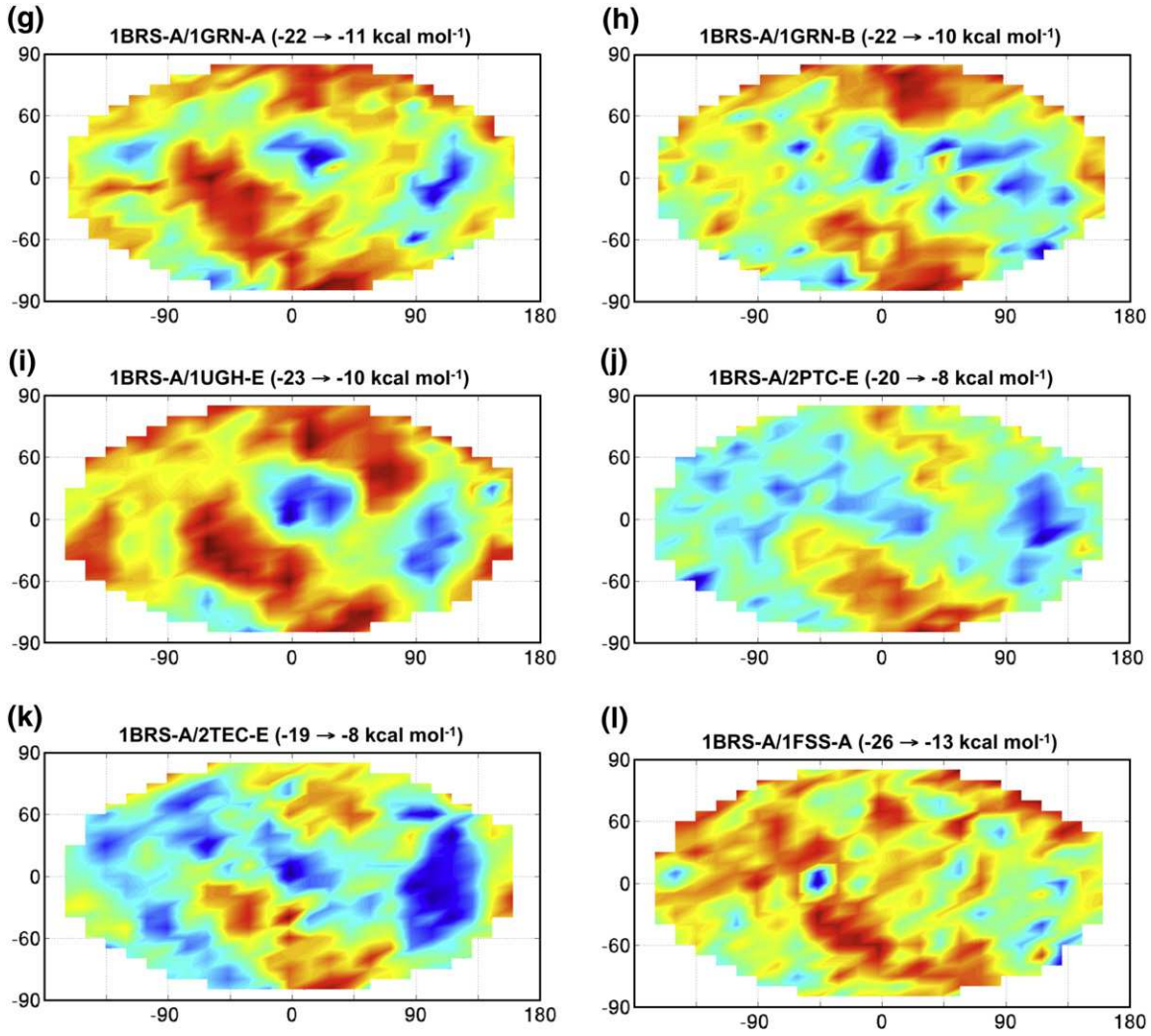


Fig. 2. Energy maps resulting from the docking of 12 different proteins on barnase with the Euler angles θ and ϕ along the vertical and horizontal axis, respectively. For each map, the experimental binding site of barnase (in its complexed form with barstar) is located at the center. Ligand protein: (a) Barstar (experimental partner); (b) PTI; (c) fasciculin; (d) Eglin C; (e) UDGI, (f) Barnase; (g) CDC42 GAP; (h) CDC42 GTPase; (i) Uracil-DNA glycosylase; (j) β-trypsin; (k) thermitase; and (l) acetylcholinesterase. Ligand proteins (b) to (l) are ordered by increasing size. The interaction energy range is indicated in each case. Blue and red areas correspond to the most negative and the least negative energies, respectively.

energies) have a role in the successful identification of the experimental protein partners. This can be seen in Fig. 5b, where the matrix shown is based purely on the normalized FIR values describing the fraction of experimentally identified interface residues that actually occur at the interaction surface of each pair. Although this is clearly an important factor, it still fails to identify three protein pairs correctly and shows a less sharp distinction between correct and incorrect partners than the NII values in Fig. 5a.

We can conclude, at least for this small test set of proteins, that if we know which residues form the binding interface of a protein, we can use our existing docking method and interaction energy to successfully identify the correct complexes and reduce the search time, since it will no longer be necessary to look at conformations that do not involve significant fractions of the correct interface residues. However, this implies identifying the "correct" interface resi-

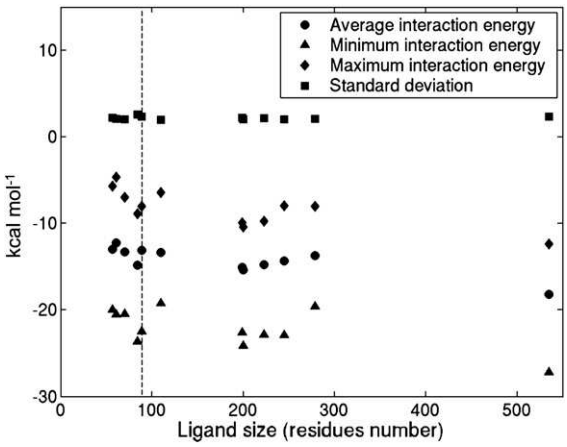


Fig. 3. Statistical data concerning the energy maps obtained for barnase. The vertical broken line crosses points corresponding to the experimental partner (barstar).

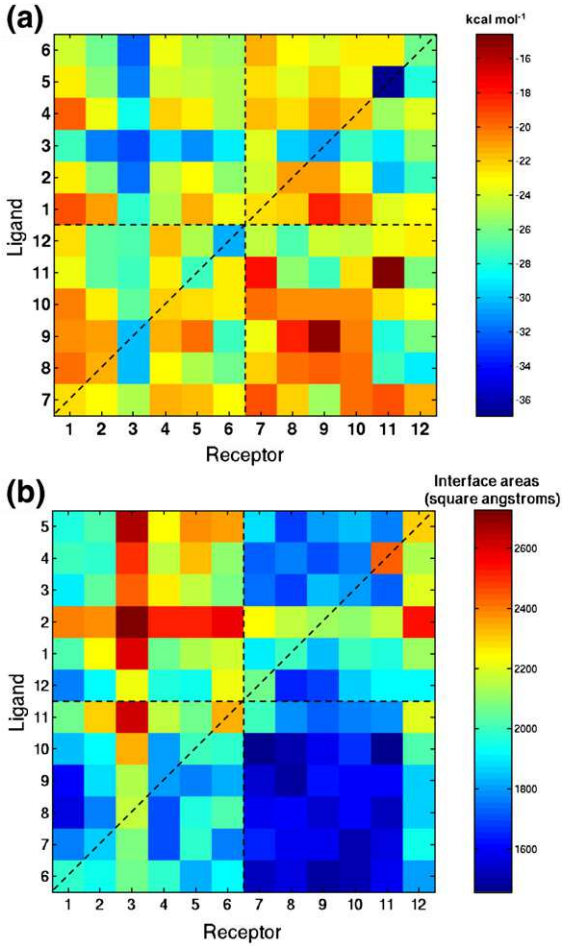


Fig. 4. CC-D matrices for the 12 proteins dataset, each protein is numbered as follows: 1, 1BRS-A; 2, 2PTC-E; 3, 1FSS-A; 4, 2TEC-E; 5, 1UGH-E; 6, 1GRN-A; 7, 1BRS-D; 8, 2PTC-I; 9, 1FSS-B; 10, 2TEC-I; 11, 1UGH-I; and 12, 1GRN-B. The receptor columns and ligand rows are ordered so that experimentally observed complexes lie on the trailing diagonal (indicated by the dashed line). Interactions between the largest proteins of each pair occur in the upper left-hand quadrant of the matrix and those between the smallest proteins in the lower right-hand quadrant. a, Minimum interaction energy matrix; b, maximum interface area matrix.

dues. This was an easy task for our trial set of proteins because we knew the structure of the experimentally identified complexes. In general, this will obviously not be the case. As we discuss below, identifying binding interfaces is an active area of research and a wide variety of methods based on both sequence analysis and physical parameters exist. These methods could be used but, on the basis of the observations for barnase described above, we can also ask whether cross-docking itself could help to identify the correct binding interfaces.

Cross-docking to identify the good interface

Early studies have suggested that docking the wrong partners together can nevertheless point to the correct interaction surfaces. We can cite, for example, the studies conducted by Fernández-Recio *et al.*, who performed docking simulations of a non-native ligand (lysozyme) with three protein receptors (chymotrypsin, cytochrome *f* and UDG), and observed an accumulation of the docking solutions around the experimental location of the native ligand in chymotrypsin.[22] In order to analyze the tendency of our docking to identify complex conformations involving the correct binding interfaces of each partner, we can define a simple interface propensity for residue i of protein P_2 in the complex P_1–P_2 as:

$$\frac{N_{\text{int},P_1P_2}(i)}{N_{conf,P_1P_2}}$$

where $N_{\text{conf},P1P2}$ is the number of orientations of P_2 tested at every surface point on the receptor P_1 (which depends on the size of P_2) and $N_{\text{int},P1P2}(i)$ is the number of these conformations where residue i

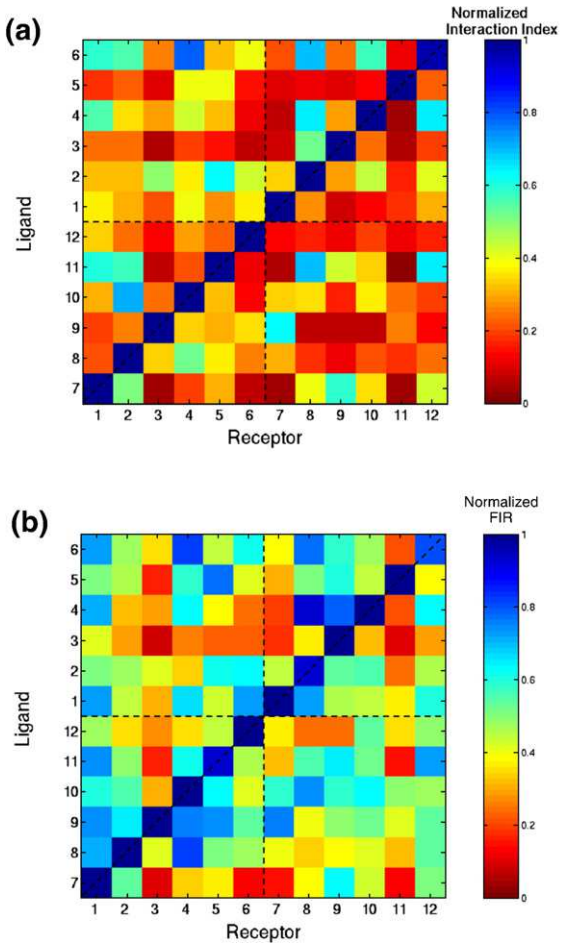


Fig. 5. (a) Normalized interaction index (NII) matrix for the 12 proteins dataset. The NII index estimates the quality of a protein–protein interaction based on correct residues appearing at the interface combined with good interaction energies. Low NII values indicate poor interaction between two proteins, while NII = 1 denotes the best possible complex. The matrix is ordered as in Fig. 4, with the experimental complexes lying on the trailing diagonal indicated by the dashed line. (b) Normalized fraction of interface residues (FIR) for the 12 proteins dataset.

belongs to the binding interface (defined again by at least a 10% decrease in its accessible surface area compared to the isolated protein P_2).[22]

The disadvantage of this value is that is that it counts all conformations of the binary complex equally, whether their interaction energy is good or bad. In order to correct this, we introduced a Bolzmann weighting that favors conformations with the most negative interaction energies:

$$IP_{P_1P_2}(i) = \frac{\sum_{j \in N_{\text{int},P_1P_2(i)}} \exp\left(-\frac{E(j)-E_0}{RT}\right)}{\sum_{j \in N_{pos,P_1P_2}} \exp\left(-\frac{E(j)-E_0}{RT}\right)}$$

where $E(j)$ is the interaction energy in conformation j, E_0 is the lowest interaction energy obtained for the P_1–P_2 complex, T is the temperature (300 K), and R is the gas constant.

As in the preceding section, this index can be normalized by using the difference between the observed value for residue i and the average over all surface residues, divided by the maximal range of IP for this particular complex:

$$NIP_{P_1P_2}(i) = \frac{IP_{P_1P_2}(i) - \langle IP_{P_iP_2}\rangle_{i \in Surf}}{\max(IP_{P_1P_2})_{i \in Surf} - \langle IP_{P_iP_2}\rangle_{i \in Surf}}.$$

NIP can then be positive, indicating that residue i is favored and occurs at the P_1–P_2 interface more commonly than would be expected statistically, or negative, indicating that it is disfavored. We then used NIP as a parameter for the prediction of protein binding sites, dividing the residues into two groups: $NIP \geq 0$ predicted as belonging to the binding interface; $NIP < 0$ predicted as not belonging to the binding interface.

If we simply group together the NIP results for all residues in a given group of complexes, we can use the classical notions of sensitivity (Sen.) and specificity (Spe.) to evaluate the usefulness of NIP for identifying interface residues. Sensitivity is defined as the number of interface residues predicted correctly (true positives) divided by the total number of experimentally defined interface residues. Specificity is the number of non-interface surface residues (true negatives) predicted correctly divided by the total number of experimentally defined non-interface residues. Optimal predictions would have both sensitivity and specificity equal to unity. If this cannot be achieved, the best compromise is obtained by minimizing the error function:

$$Err. = \sqrt{(1 - Sen.)^2 + (1 + Spe.)^2}$$

Figure 6a illustrates this error function based on NIP predictions made in two ways: first, using only experimentally observed complexes (Simple Docking or SD — broken line), and second, using all possible complexes (Complete Cross-Docking or CC-D — continuous line). The diagonal in this figure corresponds to random predictions. If we choose a cutoff value of NIP (–0.04) corresponding to the closest approach to the origin in Fig. 6(a), the simple docking allows us to select 18% of the surface residues with a sensitivity of 81% and a specificity of 91% for the true interface residues. For cross-docking, the test now takes into account artificial interfaces generated between proteins that do not truly interact. In this case, the optimal value of NIP (–0.02) enables us to select 30% of surface residues with a sensitivity of 78% and a specificity of 77% for the interface residues. Although NIP is naturally less effective in recognizing correct interface residues for CC-D than for SD, it is still surprisingly effective where only experimental complexes are studied.

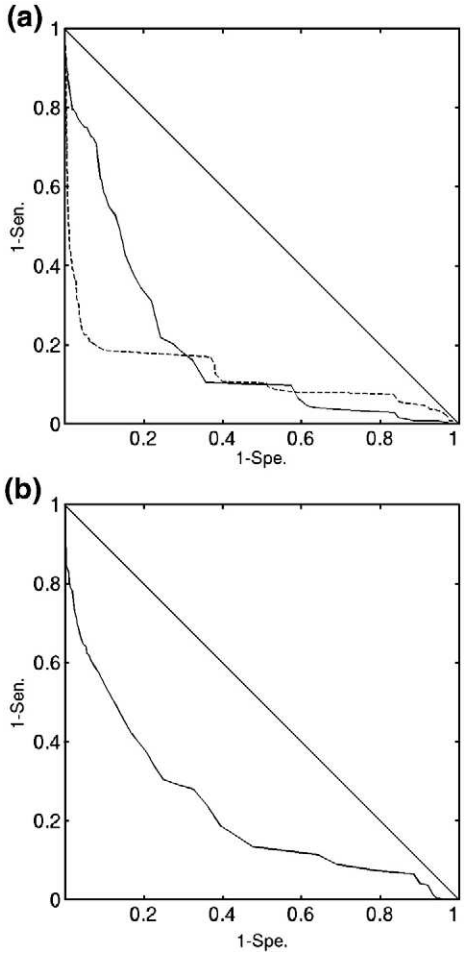


Fig. 6. Detection of interface residues using the NIP index, residues with a positive NIP values are favored and occur at the protein interface more commonly than would be expected statistically. Evolution of the sensitivity and specificity for cross-docking (continuous line) compared to simple docking (dashed line) and a random predictions (continuous diagonal line). The data in a correspond to docking using bound protein conformations, while the data in b correspond to unbound conformations.

Another way of looking at these results is shown in Fig. 7, where the selection of residues potentially belonging to a protein interface is shown as a function of the cutoff value of NIP. Figure 7a shows the results for SD (experimental complexes). The

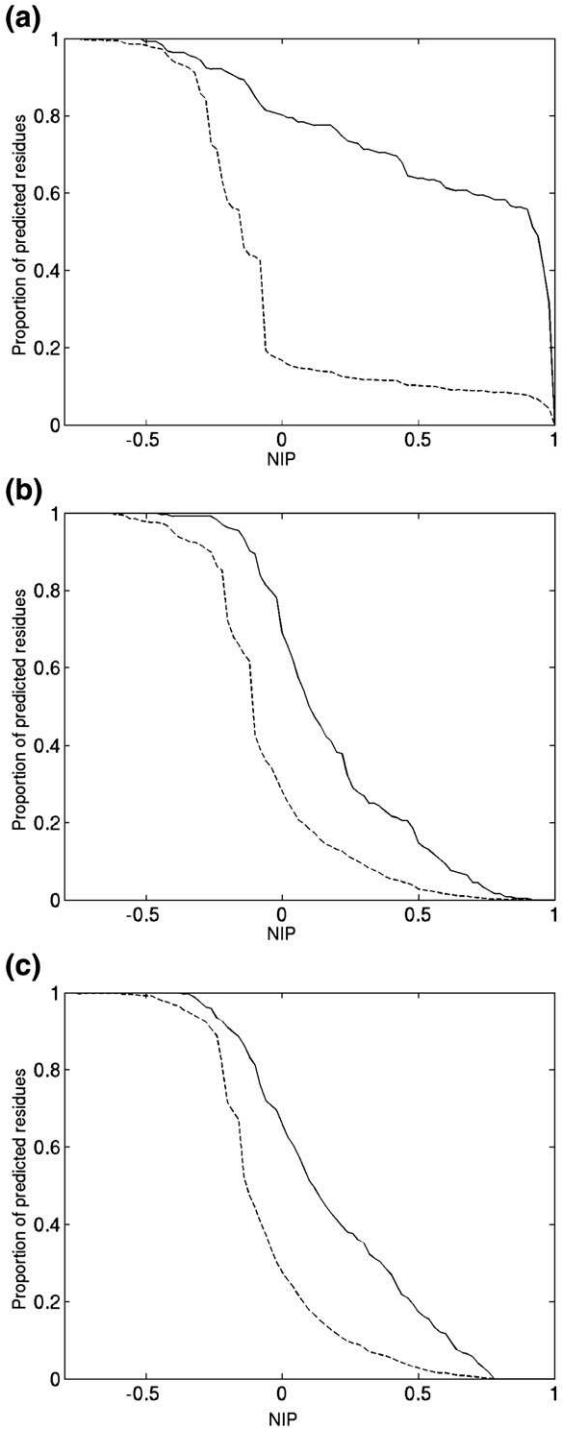


Fig. 7. Enrichment of interface residues using the NIP index shown by comparing the fraction of true interface residues detected (continuous line) with the fraction of surface residues (broken line) as a function of the NIP cutoff. The data in a and b correspond to simple docking and to cross-docking (see the text), respectively, using bound protein conformations, while the data in c corresponds to cross-docking using unbound protein conformations.

difference between the dashed line (fraction of all residues selected) and the continuous line (fraction of interface residues selected) shows the enrichment of the sample due to using the NIP criteria. When we compare with the CC-D trial in Fig. 7b, we see that the enrichment is reduced, but still significant.

Importance of protein conformational changes upon binding

Although forming the protein complexes belonging to our test set (with one exception) does not induce major conformational changes in the constituent proteins, the small backbone movements and rearrangements of side chains can be expected to have some effect on the docking and the interface predictions we carry out here. We consequently repeated the work described above using the unbound forms of the 12 proteins. As expected, the results are less satisfying. The interaction index *NII* is still a useful guide to which proteins actually form complexes, since we can identify four complexes correctly (those involving thermitase, barstar, fasciculin and CDC42 GAP as the receptor protein). On average, the experimentally identified complexes also have higher *NII* values (0.66) than for the full cross-docking set of interactions (0.43). The decreased quality of the predictions in this case is mainly due to the poorer fit of unbound structures, since the average value of *FIR* is 0.64 for CC-D with the bound structures, *versus* only 0.48 for unbound structures. At the same time, the average interaction energies weaken by only 1.7 kcal mol^{-1}.

Concerning the detection of interface residues using unbound conformations, the *NIP* criterion is surprisingly good (see Figs. 6b and 7c). If we consider the CC-D results, we find that for an optimal value of *NIP* (–0.02) we select 30% of surface residues, with a sensitivity of 70% and a specificity of 75% for interface residues. These results are only slightly worse than those found using bound structures (see the preceding section).

Influence of the type of complex considered

Five out of the six complexes in the test set described above belong to the enzyme–inhibitor category (***1GRN*** is classified as "other" in the protein benchmark[35]). Are the results that we have presented applicable to other types of complex? To begin to answer this question, we have selected eight additional proteins forming four experimentally identified complexes. These complexes also belong to the rigid-body docking section of the docking benchmark,[35] but not to the enzyme–inhibitor category. The complexes studied were the following: FAB D3H44/tissue factor (***1JPS***[39]); camel VHH/pancreatic α-amylase (***1KXQ***[40]); RAC GTPase/Pseudomonas toxin GAP domain (***1HE1***[41]); actin/vitamin D binding protein (***1KXP***[42]). In discussing these new complexes, we include the results on CDC42 GTPase/CDC42 GAP (***1GRN***), which we recall is an enzyme–activator complex that undergoes significant backbone deformation upon complex formation. We term this set TSII and we will contrast the results obtained with those already seen for the five

enzyme–inhibitor complexes already discussed (TSI), that is, ***1BRS***, ***2PTC***, ***1FSS***, ***2TEC*** and ***1UGH***. This choice yields two test sets involving ten proteins each.

We performed a CC-D trial on the TSII set in the most unfavorable case; i.e. using their unbound conformations and compared the results with the TSI set under the same conditions. Figure 8a shows the matrix for TSI. As stated earlier, CC-D trials with unbound protein structures are less satisfying than those with bound structures. The normalized interaction index (*NII*) nevertheless allows us to identify the correct partners in three cases (thermitase, barstar and fasciculin) and, on average, the experimental complexes have larger *NII* values (0.65) than the overall average for the 10 × 10 CC-D interactions (0.45).

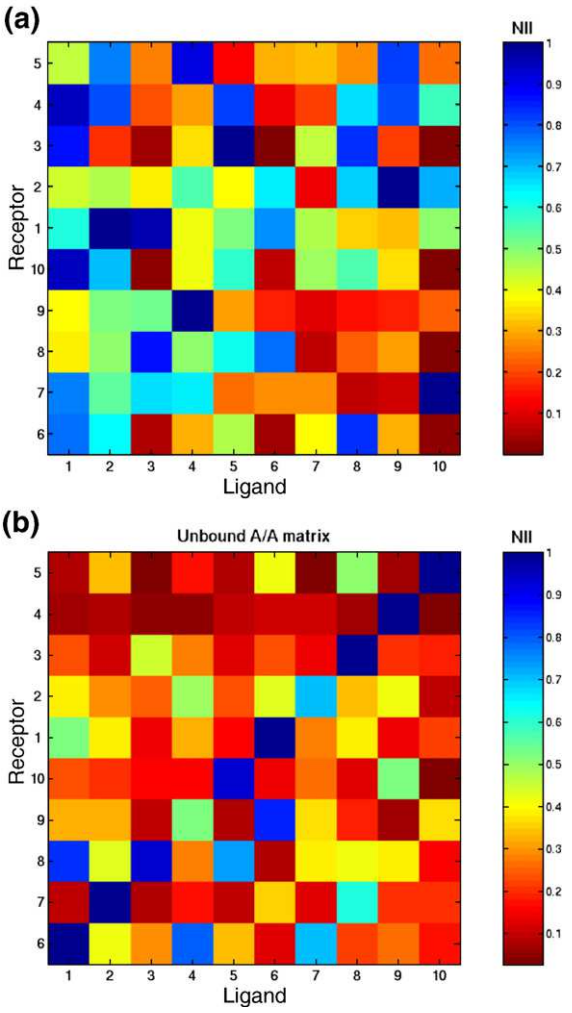


Fig. 8. Normalized interaction index (NII) matrices obtained after cross-docking calculations on the TSI (enzyme/inhibitor) and new TSII sets, using unbound protein structures in both cases. The NII index estimates the quality of a protein–protein interaction based on correct residues appearing at the interface combined with good interaction energies. Low NII values indicate poor interaction between two proteins, while NII = 1 denotes the best possible complex. (a) TSI enzyme/inhibitor set. Each protein (indicated by its PDB code and chain ID) is numbered as follows: 1, 1BRS-A; 2, 2PTC-E; 3, 1FSS-A 4; 2, TEC-E; 5,1UGH-E; 6, 1BRS-D; 7, 2PTC-I; 8, 1FSS-B; 9, 2TEC-I; 10, 1UGH-I. (b) The new TSII set. Each protein (indicated by its PDB code and chain ID) is numbered as follows: 1, 1GRN-A; 2, 1HE1-C; 3, 1JPS-HL; 4, 1KXP-A; 5, 1KXQ-H; 6, 1GRN-B; 7, 1HE1-A; 8, 1JPS-T; 9, 1KXP-D; and 10, 1KXQ-A. In a and b, the rows and columns are ordered so that experimentally observed complexes lie on the trailing diagonal of the matrix.

The results for the CC-D trial on the new TSII set are still more encouraging, despite the use of unbound structures. As we can see in Fig. 8b, the trailing diagonal of the matrix (corresponding to the experimentally identified complexes) is now clearly distinguished from the incorrect off-diagonal complexes. Using *NII*, we can correctly identify the experimental interaction partner of eight proteins out of ten (CDC42 GTPase, CDC42 GAP, RAC GTPase, FAB D3H44, tissue factor, vitamin D binding protein, camel VHH and pancreatic α-amylase) and, on average, the experimental complexes present an *NII* value of 0.91, *versus* an overall average of 0.32 for the 10 × 10 CC-D interactions.

Predicting interface residues

Our results have stressed the importance of identifying the interface residues in order to identify correct protein complexes. CC-D has been shown to help in obtaining this data, but many other methods exist for identifying protein-binding interfaces, based on physical criteria or on sequence analyses (see the recent review by Zhou and Qin[43]) and it is worth comparing such approaches with the results from CC-D. We have presently tested two alternative approaches that are easily accessible via web servers:

(a) Cons-PPISP, which uses a PSI-Blast sequence profile and solvent accessibility as input to a neural network†.[44]
(b) Promate, which uses a Bayesian method based on properties such as secondary structure, atom distribution, amino acid pairing and sequence conservation‡.[45]

We contrasted the results of these two approaches (using default parameters) with our normalized interface propensity (*NIP*) criterion based purely on CC-D trials, defining interface residues as those with a positive *NIP* values. The results are summarized in Table 2 by comparing the statistical quality of interface residue detection, and the average fraction of interface residues (*FIR*) and *NII* values after CC-D trials. All these results refer to the TSI protein set (enzyme/inhibitor complexes, see above) and bound protein structures. We note that both cons-PPISP and Promate predictions are more selective than those obtained with the *NIP* value. They both

† http://pipe.scs.fsu.edu/ppisp.html
‡ http://bioportal.weizmann.ac.il/promate

Table 2. Using web servers or the NIP index for the prediction of interface residues

Interface prediction method	Experimental data	Cons-PPISP	Promate	Positive NIP
Coverage[a] (%)	100	10	8	27
Sensitivity[b] (%)	100	40	25	69
Specificity[c] (%)	100	94	95	79
Accuracy[d] (%)	100	44	38	31
<FIR>	0.66	0.32	0.24	0.26
<NII>[e]	0.35	0.34	0.30	0.34
$\langle NII \rangle_{expt}$	1.0	0.40	0.33	0.33
NScore[f]	0.62	0.79	0.79	0.82

[a] Fraction of residues predicted as interface residues (over a complete set of 1721 residues).
[b] Fraction of experimental interface residues that are correctly predicted as such (over a set of 206 residues).
[c] Fraction of experimental non-interface residues correctly predicted as such.
[d] Number of true positives divided by the total number of predicted residues.
[e] <NII> is the average value of NII over the full cross-docking set of interactions, while <NII>expt is its average value on the experimentally identified complexes alone.
[f] The score of a cross-docking matrix is obtained by reordering its rows and columns (while maintaining experimental pairs on the diagonal) so that the moment $S = \sum_{i=1}^{Nprot} \sum_{j=1}^{Nprot} d_{ij} NII_{ij}$ is minimized. *N*prot is the number of proteins in the test set, and $d_{ij} = \frac{|i-j|}{\sqrt{2}}$ is the distance of the i-j protein pair to the diagonal. We then define the normalized score $NScore = \frac{S}{S_{max} \times \langle NII \rangle}$, where S_{max} is the score for a matrix of the corresponding dimension (e.g., 233.3 for 10×10) and containing 1 in all entries.

exhibit a very low sensitivity (under 40%), unlike *NIP* with a value of 69%. In contrast, both web server approaches have higher specificity and accuracy values. Unfortunately, none of the three methods is comparable in performance to using experimentally defined interfaces, when it comes to comparing the *NII* scores for all proteins pairs. In all three cases, the discrepancy between the average *NII* for the whole CC-D trial set and for the true experimental complexes almost disappears, rendering the identification of experimental interaction partners impossible. This failure is due mainly to the poor estimation of *FIR* whose average value on the whole set drops from 0.66, when using the experimental data, to 0.24, 0.26 and 0.32 when using the predictions of Promate, *NIP*, and cons-PPISP respectively.

We can get an overall view of how these different methods would work in CC-D trials by calculating the dominance of the trailing diagonal in the TSI 10 × 10 matrix. This can be done using a score *S* equivalent to the moment of the off-diagonal elements:

$$S = \sum_{i=1}^{Nprot} \sum_{j=1}^{Nprot} d_{ij} NII_{ij}$$

where d_{ij} is the distance of element *ij* from the trailing diagonal, $d_{ij} = \frac{|i-j|}{\sqrt{2}}$, and NII_{ij} is the normalized interaction index defined above. Note that *S* is minimized for each matrix by reordering the rows and columns, keeping the experimentally identified protein partners along the trailing diagonal, and can be normalized as:

$$NScore = \frac{S}{S_{max} \times \langle NII \rangle}$$

where S_{max} is the score for a matrix of the appropriate dimension with ones in all entries (233.3 for 10×10). Smaller values of *NScore* quantify the dominance of the correct protein partners (along the trailing diagonal) compared to all other possible pairs. As can be seen in Table 2, the cons-PPISP and Promate approaches perform equally in identifying the correct complexes (*NScore* = 0.79), and slightly better than our *NIP* index (*NScore* = 0.82). However, not surprisingly, none of these methods is as good as using experimentally identified interface residues (*NScore* = 0.62).

Discussion

The results of CC-D on two small protein datasets have shown that the rigid-body docking method we have used is good at determining the conformation of experimentally identified binary complexes, but cannot distinguish between correct and incorrect complexes. An analysis of the interaction interfaces suggests that solving this problem requires determining the correct interfaces before attempting to dock, which is not an easy task.

Previous studies have shown that the protein–protein interface composition, unlike that of protein–ligand interfaces,[46] does not differ significantly from the rest of the protein surface.[47–49] It is worth pointing out, however, that our *NIP* index, which is based purely on the likelihood of a given residue appearing in the optimal interface with another protein (irrespective of whether this protein pair actually forms a complex), also provides useful information. We find that this information is almost as successful in defining interface residues as other methods based on physical properties or sequence analysis, such as those reviewed by Zhou and Qin.[43] In tests for our TSI dataset, using bound protein conformations, cons-PPISP[44] and Promate[45] did only slightly better than NIP in favoring experimentally-defined complexes, although none of these methods was as good as using experimentally defined interfaces.

Further difficulties are likely to be encountered for proteins with several distinct interaction surfaces created by multiple interactions. The second edition of the docking benchmark of Mintseris *et al.*[35] contains many such cases (with complexes ranging from trimers to hexamers) and merits further studies.

The final problem is linked to changes in conformation coupled with the formation of a complex. Unsurprisingly, the use of unbound structures for cross-docking decreased the efficiency with which we could predict the correct partners.[30] When no major backbone rearrangement occurs (as for most of the proteins in our TSI and TSII test sets), this problem

could probably be overcome by taking the optimization of side chain conformation into account.

Conclusions

We have addressed the problem of whether a rigid-body, coarse-grain docking method can detect which proteins should form binary complexes within two test sets of proteins. The results demonstrate that although docking the correct partners leads to conformations close to those observed experimentally, the energy score used cannot discriminate between these protein pairs and other incorrect combinations.

However, we show that if the interface residues of each protein involved in the correct complexes can be identified, then it is possible to eliminate incorrect partners due to the fact that they preferentially form complexes that do not involve these residues or that they lead to poor interaction scores when these residues lie at the interface. We have developed an index on this basis (using experimental information on the interfaces) that correctly identifies all experimental complexes when docking is carried out with bound protein conformations and correctly identifies a subset of complexes when unbound conformations are used. This means that the problem we posed can be redefined as the need to identify the correct interaction interfaces.

We have shown also that docking incorrect protein partners leads to complexes that tend to use the correct interaction interfaces and that this information alone is almost as successful as other sequence or physical property-based approaches in defining these interfaces. However, none of the methods tested is sufficiently accurate to currently solve the partner identification problem.

These results certainly need to be verified on larger and more diverse protein sets. Hopefully, more extensive complete cross-docking will improve the identification of interaction interfaces and could open the route for some iterative improvement of this information on the basis of the structural and energetic quality of the interfaces formed. We will also attempt to allow for structural adaptation, at least of amino acid side chains, using a multi-copy approach in the spirit of that already used by Bastard *et al.*[50] To counterbalance the necessary increase in calculation time, it should be possible to restrict the exploration of the receptor surface to the areas surrounding binding sites predicted by the methods discussed above. We are also studying other methods based on evolutionary sequence information[51–54] in order to try and improve the definition of these areas.

This study can be seen as a "pre-docking" step, in that our aim is to find interacting partners, rather than to accurately identify the conformation of binary complexes. Although this step has not received much attention until now, it is becoming more important as interest shifts to understanding interactions on the proteomic scale. Our results suggest clearly that this step differs significantly from docking known interaction partners and needs new approaches, notably for identifying interaction interfaces.

Materials and Methods

In this section, we describe the MAXDo (Molecular Association via Cross Docking) algorithm that was developed for CC-D studies. Since complete cross-docking involves a much larger number of calculations than simple docking, we chose a rigid-body docking approach with a reduced protein model in order to make rapid conformational searches. All simulations were initially performed using the bound conformations of the proteins, but we also consider their unbound conformations (both being extracted from the Protein Data Bank[55]).

Reduced protein representation

We have used a coarse-grain protein model developed by Zacharias,[56] where each amino acid is represented by one pseudoatom located at the Cα position, and either one or two pseudoatoms representing the side chain (with the exception of Gly). Ala, Ser, Thr, Val, Leu, Ile, Asn, Asp, and Cys have a single pseudoatom located at the geometrical center of the side-chain heavy atoms. For the remaining amino acids, a first pseudoatom is located midway between the Cβ and Cγ atoms, while the second is placed at the geometrical center of the remaining side-chain heavy atoms. This description, which allows different amino acids to be distinguished from one another, has already proved useful in protein–protein docking[50,56,57] and protein mechanics studies.[58,59]

Interactions between the pseudoatoms of the Zacharias representation are treated using a soft LJ-type potential with appropriately adjusted parameters for each type of side chain (see Table 1 in Ref.[42]). In the case of charged side chains, electrostatic interactions between net point charges located on the second side chain pseudoatom were calculated by using a distance-dependent dielectric constant $\varepsilon = 15r$, leading to the following equation for the

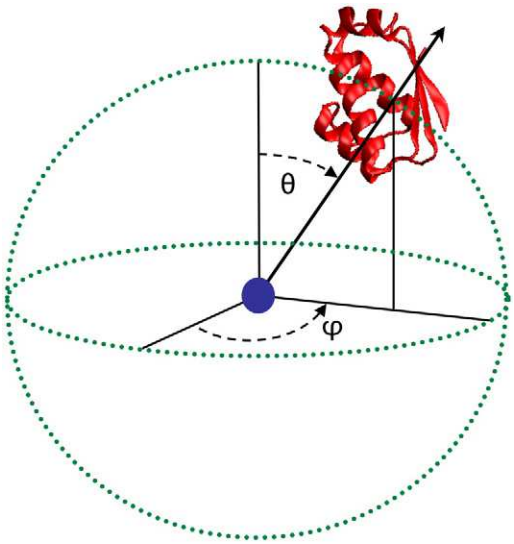


Fig. 9. Summary of the docking algorithm. For each of the starting positions, defined by the Euler angles θ and ϕ, evenly spaced around the receptor protein (blue point), the ligand protein (in red) can change its orientation and its distance from the receptor during the energy minimization.

interaction energy of the pseudoatom pair i,j at distance r_{ij}:

$$E_{ij} = \left(\frac{B_{ij}}{r_{ij}^8} - \frac{C_{ij}}{r_{ij}^6}\right) + \frac{q_i q_j}{15 r_{ij}^2}$$

where B_{ij} and C_{ij} are the repulsive and attractive LJ-type parameters, respectively, and q_i and $q_j$are the charges of the pseudoatoms i and j.

Systematic docking simulations

Our systematic docking algorithm (Fig. 9) was derived from the ATTRACT protocol of Zacharias[56] and uses a multiple energy minimization scheme. For each pair of proteins, the first molecule (called receptor) was fixed in space, while the second (termed the ligand protein) was used as a probe and placed at multiple positions on the surface of the receptor. The distance of the probe from the receptor was chosen so that no pair of probe-receptor pseudoatoms came closer than 6 Å. Starting probe positions were randomly created around the receptor surface with a density of one position per 10 Å^2, and for each starting position, 210 different ligand orientations were generated, resulting in a total number of start configurations ranging from 95,000 to 450,000 depending on the size of the receptor.

During each energy minimization, the ligand protein was kept at a given location over the surface of the receptor protein by using a harmonic restraint to maintain its center of mass on a vector passing through the center of mass of the receptor protein. The direction of this vector was defined by two Euler angles θ and φ, (where θ = φ, = 0° was chosen to pass through the center of the binding interface of the receptor protein) as shown in Fig. 9. By varying the Euler angles from 0°→360° and 0°→180°, respectively, it was possible to sample interactions evenly over the complete surface of the receptor and to represent its binding potential using 2D energy maps (each point corresponding to the best ligand orientation for the chosen θ/φ pair).

Computational implementation

Each energy minimization for a pair of interacting proteins takes typically 15 s on a single 2 GHz processor. As stated above, 100,000 ~ 450,000 minimizations are required to probe all possible interaction conformations, depending on the size of the interacting proteins. This would require many days of computation on a single processor. Happily, each minimization is independent of the others and this problem therefore belongs to the so-called "embarrassingly parallel" category and is perfectly adapted to petaflop machines with very large numbers of processors, or to grid calculations. In the present case, the CC-D trials have been performed on the DECRYTHON university grid§, where the docking of a single protein pair took between 5 h and two days (depending on the size of the proteins and the grid availability). The MAXDo program is currently being refined and it is expected that it will be possible to gain a factor of 4–5. Larger CC-D trials will be carried out using the World Community Grid‖ whose massive computational power will enable us to scan CC-D sets involving thousands of proteins.

§ www.decrypthon.fr
‖ www.worldcommunitygrid.org

Acknowledgements

This work was carried out in the framework of the DECRYPTHON Project, set up by the CNRS (Centre National de la Recherche Scientifique), the AFM (French Muscular Distrophy Association) and IBM. The cross-docking simulations were performed on the Decrypthon University Grid, and we thank Raphaël Bolze for adapting our docking program for use on this grid. S.S.-M. thanks the AFM for one year of post-doctoral funding during which part of this work was carried out.

References

1. Ellis, R. J. & Minton, A. P. (2003). Cell biology - join the crowd. *Nature*, **425**, 27–28.
2. Minton, A. P. (2000). Implications of macromolecular crowding for protein assembly. *Curr. Opin. Struct. Biol.* **10**, 34–39.
3. Deeds, E. J., Ashenberg, O., Gerardin, J. & Shakhnovich, E. I. (2007). Robust protein-protein interactions in crowded cellular environments. *Proc. Natl Acad. Sci. USA*, **104**, 14952–14957.
4. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17**, 1030–1032.
5. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R. *et al.* (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.
6. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
7. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L. *et al.* (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.
8. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A. *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
9. Xenarios, L. & Eisenberg, D. (2001). Protein interaction databases. Curr. *Opin. Biotechnol.* **12**, 334–339.
10. Shoemaker, B. A. & Panchenko, A. R. (2007). Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *Plos Comput. Biol.* **3**, 337–344.
11. Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F. F., Pawson, T. & Hogue, C. W. V. (2001). BIND - the biomolecular interaction network database. *Nucleic Acids Res.* **29**, 242–245.
12. Shoemaker, B. A. & Panchenko, A. R. (2007). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *Plos Comput. Biol.* **3**, 595–601.
13. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. (2002). Protein interactions - Two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics*, **1**, 349–356.
14. Levinthal, C., Wodak, S. J., Kahn, P. & Dadivanian, A. K. (1975). Hemoglobin interaction in sickle-cell fibers .1. Theoretical approaches to molecular contacts. *Proc. Natl Acad. Sci. USA*, **72**, 1330–1334.

15. Wodak, S. J. & Janin, J. (1978). Computer analysis of protein–protein interaction. *J. Mol. Biol.* **124**, 323–342.
16. Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
17. Mendez, R., Leplae, R., Lensink, M. F. & Wodak, S. J. (2005). Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins: Struct. Funct. Bioinform.* **60**, 150–169.
18. Carter, P., Lesk, V. I., Islam, S. A. & Sternberg, M. J. E. (2005). *Protein-protein docking using*, **3**, 281–288.
19. Bahadur, R. P. & Zacharias, M. (2008). The interface of protein-protein complexes: Analysis of contacts and prediction of interactions. *Cell. Mol. Life Sci.* **65**, 1059–1072.
20. Lesk, V. I. & Sternberg, M. J. E. (2008). 3D-Garden: a system for modelling proteinprotein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm. *Bioinformatics*, **24**, 1137–1144.
21. Lensink, M. F., Mendez, R. & Wodak, S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins: Struct. Funct. Bioinform.* **69**, 704–718.
22. Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2004). Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.* **335**, 843–865.
23. Elcock, A. H. (2002). Atomistic Simulations of competition between substrates binding to an enzyme. *Biophys. J.* **82**, 2326–2332.
24. Macchiarulo, A., Nobeli, I. & Thornton, J. M. (2004). Ligand selectivity and competition between enzymes in silico. *Nature Biotechnol.* **22**, 1039–1045.
25. Yang, J. M. & Chen, C. C. (2004). GEMDOCK: A generic evolutionary method for molecular docking. *Proteins: Struct. Funct. Bioinform.* **55**, 288–304.
26. Sotriffer, C. A. & Dramburg, I. (2005). "In situ cross-docking" to simultaneously address multiple targets. *J. Med. Chem.* **48**, 3122–3125.
27. Meiler, J. & Baker, D. (2006). ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins: Struct. Funct. Bioinform.* **65**, 538–548.
28. Bottegoni, G., Kufareva, I., Totrov, M. & Abagyan, R. (2008). A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). *J. Comput.Aided Mol. Des.* **22**, 311–325.
29. Claussen, H., Buning, C., Rarey, M. & Lengauer, T. (2001). FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **308**, 377–395.
30. Smith, G. R., Sternberg, M. J. E. & Bates, P. A. (2005). The relationship between the flexibility of proteins and their conformational states on forming protein–protein complexes with an application to protein-protein docking. *J. Mol. Biol.* **347**, 1077–1101.
31. Zhao, Y. & Sanner, M. F. (2007). FLIPDock: Docking flexible ligands into flexible receptors. *Proteins: Struct. Funct. Bioinform.* **68**, 726–737.
32. Krol, M., Chaleil, R. A. G., Tournier, A. T. & Bates, P. A. (2007). Implicit flexibility in protein docking: Cross-docking and local refinement. *Proteins: Struct. Funct. Bioinform.* **69**, 750–757.
33. Lee, H. S., Choi, J., Kufareva, I., Abagyan, R., Filikov, A., Yang, Y. & Yoon, S. (2008). Optimization of high throughput virtual screening by combining shape-matching and docking methods. *J. Chem. Informat. Model.* **48**, 489–497.
34. Chen, R., Mintseris, J., Janin, J. & Weng, Z. P. (2003). A protein-protein docking benchmark. *Proteins: Struct. Funct. Genet.* **52**, 88–91.
35. Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J. & Weng, Z. P. (2005). Protein-protein docking benchmark 2.0: An update. *Proteins: Struct. Funct. Bioinform.* **60**, 214–216.
36. Brooijmans, N., Sharp, K. A. & Kuntz, I. D. (2002). Stability of macromolecular complexes. *Proteins: Struct. Funct. Genet.* **48**, 645–653.
37. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2004). A dissection of specific and non-specific protein – protein interfaces. *J. Mol. Biol.* **336**, 943–955.
38. Hubbard, S. J. (1992). *ACCESS: a program for calculating accessibilities, Department of Biochemistry and Molecular Biology.* University College of London, .
39. Faelber, K., Kirchhofer, D., Presta, L., Kelley, R. F. & Muller, Y. A. (2001). The 1.85 Å resolution crystal structures of tissue factor in complex with humanized Fab D3h44 and of free humanized Fab D3h44: Revisiting the solvation of antigen combining sites. *J. Mol. Biol.* **313**, 83–97.
40. Desmyter, A., Spinelli, S., Payan, F., Lauwereys, M., Wyns, L., Muyldermans, S. & Cambillau, C. (2002). Three Camelid VHH domains in complex with porcine pancreatic alpha-amylase - Inhibition and versatility of binding topology. *J. Biol. Chem.* **277**, 23645–23650.
41. Wurtele, M., Wolf, E., Pederson, K. J., Buchwald, G., Ahmadian, M. R., Barbieri, J. T. & Wittinghofer, A. (2001). How the Pseudomonas aeruginosa ExoS toxin downregulates Rac. *Nature Struct. Biol.* **8**, 23–26.
42. Otterbein, L. R., Cosio, C., Graceffa, P. & Dominguez, R. (2002). Crystal structures of the vitamin D-binding protein and its complex with actin: Structural basis of the actin-scavenger system. *Proc. Natl. Acad. of Sci. USA*, **99**, 8003–8008.
43. Zhou, H. X. & Qin, S. B. (2007). Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, **23**, 2203–2209.
44. Chen, H. L. & Zhou, H. X. (2005). Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. *Proteins: Struct. Funct. Bioinform.* **61**, 21–35.
45. Neuvirth, H., Raz, R. & Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* **338**, 181–199.
46. Burgoyne, N. J. & Jackson, R. M. (2006). Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics*, **22**, 1335–1342.
47. Chothia, C. & Janin, J. (1975). Principles of protein-protein recognition. *Nature*, **256**, 705–708.
48. Janin, J. & Chothia, C. (1990). The structure of protein-protein recognition sites. *J. Biol. Chem.* **265**, 16027–16030.
49. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
50. Bastard, K., Prevost, C. & Zacharias, M. (2006). Accounting for loop flexibility during protein-protein docking. *Proteins: Struct. Funct. Bioinform.* **62**, 956–969.
51. Dong, Q. W., Wang, X. L., Lin, L. & Guan, Y. (2007). Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. *BMC Bioinformat.* **8**.
52. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.

Functional Modes and Residue Flexibility Control the Anisotropic Response of Guanylate Kinase to Mechanical Stress

Sophie Sacquin-Mora,[†*] Olivier Delalande,[‡] and Marc Baaden[†*]
[†]Institut de Biologie Physico-Chimique, Laboratoire de Biochimie Théorique, CNRS UPR9080, Paris, France; and [‡]Centre de Biophysique Moléculaire, CNRS UPR4301, Orléans, France

ABSTRACT The coupling between the mechanical properties of enzymes and their biological activity is a well-established feature that has been the object of numerous experimental and theoretical works. In particular, recent experiments show that enzymatic function can be modulated anisotropically by mechanical stress. We study such phenomena using a method for investigating local flexibility on the residue scale that combines a reduced protein representation with Brownian dynamics simulations. We performed calculations on the enzyme guanylate kinase to study its mechanical response when submitted to anisotropic deformations. The resulting modifications of the protein's rigidity profile can be related to the changes in substrate binding affinity observed experimentally. Further analysis of the principal components of motion of the trajectories shows how the application of a mechanical constraint on the protein can disrupt its dynamics, thus leading to a decrease of the enzyme's catalytic rate. Eventually, a systematic probe of the protein surface led to the prediction of potential hotspots where the application of an external constraint would produce a large functional response both from the mechanical and dynamical points of view. Such enzyme-engineering approaches open the possibility to tune catalytic function by varying selected external forces.

INTRODUCTION

The importance of protein flexibility and dynamics for the understanding of protein function has now been clearly established (1–4). During the execution of their biological function, proteins can be subjected to forces and their mechanical properties evolved in response to fit this selection pressure. Experimentally, many techniques, such as optical and magnetic tweezers (5–7) or atomic force microscopy (8–10), make it possible to probe biomolecular mechanics directly on the single-molecule level (11). In particular, experiments with linkages other than the usual N-to-C-terminal have shown how these mechanical properties strongly depend on the loading geometry (12–15).

Although the first experiments mostly investigated the mechanical response of proteins and the sequence of unfolding events that would result from the application of a force (8), recent setups have focused more on their functional response, thus leading to the field of mechanoenzymatics (16,17). In this perspective, the mechanism of allosteric control (18,19) of an enzyme, which plays a crucial part in signaling pathways in the cell (20), can now be studied experimentally via the building of protein-DNA chimeras where a DNA molecular spring is coupled to the protein at specific locations on its surface (21–23). Through this allosteric spring probe (ASP), one can affect the static and dynamic conformations of the protein and follow its functional response to the application of an external stress (24).

From the theoretical point of view, atomic coordinate-based methods, such as constrained molecular dynamics simulations, can mimic force-extension experiments but operate on much shorter timescales and remain computationally expensive (25–29). Therefore, lower-resolution models have been widely used in recent years to study protein dynamics (30–35). These coarse-grained representations comprise the elastic network model (ENM) (36,37), which reduces the protein to a set of pseudoatoms with pairs below a given cutoff distance being linked by Gaussian springs. Despite their simplicity, these models led to many results concerning protein mechanics and dynamics (38–44). Recently, coarse-grained approaches were used to successfully model the anisotropy of the mechanical response of proteins subjected to an external force (45–47).

In this work, we used a method combining a coarse-grained protein representation and Brownian dynamics simulations. This approach was previously successfully applied to model the mechanical response of the green fluorescent protein to understand the single-molecule experiments carried out by Dietz et al. (14,47). Here, we investigated the enzyme guanylate kinase (GK), which was studied by Tseng et al. via the ASP approach (48). GK is an essential enzyme that catalyzes the transfer of a phosphate group from adenosine triphosphate (ATP) to guanosine monophosphate (GMP) (49). Upon substrate binding, GK undergoes a structural transition from the open to the closed state through a movement of the two lobes formed by the LID and GMP domains (see Fig. 1), leading to an ~1 nm conformational change (50,51). With DNA springs anchored on three different locations on the protein surface, Tseng et al. determined for each case the changes in substrate binding affinities and catalytic rate constant resulting from the directional stress exerted on the protein. They showed that the functional response strongly depends on the direction of load. Using our

Submitted July 23, 2010, and accepted for publication September 15, 2010.

*Correspondence: sacquin@ibpc.fr or baaden@smplinux.de

Editor: Nathan Andrew Baker.

0006-3495/10/11/3412/8 $2.00

doi: 10.1016/j.bpj.2010.09.026

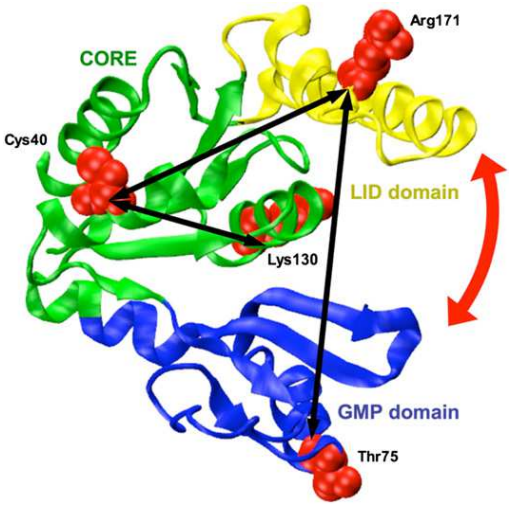


FIGURE 1 A cartoon representation of GK with the pulling directions tested experimentally by Tseng et al. (48) The color coding of the protein is according to the domain definitions of Hible et al. (67) The GMP and ATP binding sites are located at the GMP/CORE and CORE/LID interfaces, respectively. The arrow indicates the direction of the opening/closing transition, which constitutes the first mode of motion of the protein. The images in this figure and in the upper part of Fig. 5 were prepared using visual molecular dynamics (71).

molecular modeling approach, we investigated the mechanics and dynamics of GK when subjected to an external constraint and related our results to the variations of the enzymatic activity observed experimentally.

COMPUTATIONAL DETAILS

Brownian Dynamics simulations

Rigidity profile of a protein

Coarse-grained Brownian Dynamics (BD) simulations were run using a modified version of the ProPHet (probing protein heterogeneity) program (41,42), where an external mechanical constraint can be applied between two residues. In this approach, the protein is represented using an ENM. Diverging from most common coarse-grained models, where each residue is described by a single pseudoatom (52), we chose a more detailed representation (53) that involves up to three pseudoatoms per residue and enables different amino acids to be distinguished. Pseudoatoms closer than the cutoff parameter, $R_c = 9$ Å, are joined by Gaussian springs that all have identical spring constants of $\gamma = 0.42$ N m^{-1} (0.6 kcal mol^{-1} Å^{-2}). The springs are taken to be relaxed for the experimentally observed conformation of the protein, in this case the crystallographic structure of guanylate kinase from *Mycobacterium tuberculosis* in its open conformation available in the protein data bank with the code 1S4Q.

Mechanical properties are obtained from 200,000 BD steps at 300 K. The simulations are analyzed in terms of the fluctuations of the mean distance between each pseudoatom belonging to a given amino acid and the pseudoatoms belonging to the remaining residues of the protein. The inverse of these fluctuations yields an effective force constant k_i that describes the ease of moving a pseudoatom with respect to the overall protein structure:

$$k_i = \frac{3k_BT}{\langle (d_i - \langle d_i \rangle)^2 \rangle},$$

where $\langle\rangle$ denotes an average taken over the whole simulation and $d_i = \langle d_{ij} \rangle_{j^*}$ is the average distance from particle i to the other particles j in the protein (the sum over j^* implies the exclusion of the pseudoatoms belonging to residue i). The distances between the C_α pseudoatom of residue i and the C_α pseudoatoms of the adjacent residues $i - 1$ and $i + 1$ are excluded, since the corresponding distances are virtually constant. The force constant for each residue k is the average of the force constants for all its constituent pseudoatoms i. We will use the term rigidity profile to describe the ordered set of force constants for all the residues of the protein.

Applying an external constraint on the protein

Whereas in our previous work on the green fluorescent protein (47), the mechanical stress was simply modeled by applying a constant force between the C_α pseudoatoms of the corresponding residues, in this study, we chose to model the external constraint by adding to the ENM representation a supplementary spring termed the constraint spring in opposition to the structural springs resulting from the original conformation of the protein,. This way we could model more accurately the experiment of Tseng et al. (48), where DNA molecular springs of identical length (60 bp) were used at three different locations on the surface of the protein to apply a controlled mechanical stress (see Fig. 1). From the available experimental data regarding the contour length of the DNA spring (200 Å) and the amount of elastic energy that resides in the protein (54) (~1 kT), we derived for our constraint spring the parameters equilibrium length ($L_C = 150$ Å) and spring constant ($\gamma_C = 0.84$ N m^{-1}) (1.2 kcal mol^{-1} Å^{-2}). This constraint spring was added between the C_α pseudoatoms of the anchor residues of the three locations tested in the experiment, Thr^{75}/Arg^{171}, Cys^{40}/Arg^{171}, and Cys^{40}/Lys^{130} (see Fig. 1).

Principal component analysis of the coarse-grained trajectories

The BD trajectories for the protein without (relaxed protein) and with the application of an external force (protein under stress) were investigated using principal component analysis (PCA) (55–58) with tools from the Gromacs (59–61) software package. In particular, we calculated the inner product matrices of the ten first eigenvectors, which always cover >89% of the total variance of the protein (see Fig. S1,

Fig. S2, and Table S1 in the Supporting Material) of each constrained trajectory with the 10 first eigenvectors of the relaxed trajectory to assess how the mechanical constraint affects the leading modes of motion of the protein.

Systematic scan of the protein surface

To determine whether the deformations that were studied experimentally are representative of the full heterogeneity of the GK structure, we performed a more systematic study of residue-pair deformations. The selection of representative pairs is first narrowed by limiting our choice to surface residues, that is, residues with at least 5% solvent accessibility (62) (as calculated by the NACCESS program (63)), which are thus amenable to experimental study. Second, we chose residue pairs separated by at least 20 Å and 30 amino acids in the primary sequence. Last, we eliminated residue pairs that differed from already-selected pairs by fewer than five residues in the primary sequence in either of the constituent residues. This method led to the selection of 236 residue pairs, which were all tested with the constraint spring described earlier (see Fig. S3).

RESULTS

Mechanical properties of guanylate kinase

The rigidity profile of GK is represented in Fig. 2 *a*. It is worthy of remark that most of the force-constant peaks from the first half of the protein sequence correspond to residues belonging to ligand-binding sites, such as Lys^{34} and Glu^{119} for the ATP/Mg^{2+}-binding site, or Ser^{53}, Glu^{88}, and Thr^{101} for the GMP-binding site. In particular, note the peaks corresponding to Ser^{27} and Lys^{34}, two residues surrounding the flexible P-loop, a highly conserved motif that binds the β-phosphate of the ATP donor in nucleoside monophosphate kinases (64,65); and the highly rigid area on the β7-sheet, around Glu^{119}, which corresponds to residues interacting with GMP in the closed conformations of GK from *Saccharomyces cerevisiae* (50) and from *Mus musculus* (66).

The variation of the force-constant profile of the protein upon mechanical stress is represented in Fig. 2 *b*. The P-loop base and the β7-sheet are the protein segments whose mechanical properties are the most sensitive to the application of an external stress, independent of the pulling direction. The mechanical response of GK is nevertheless markedly anisotropic. Although the 40/171 and 40/130 pulling directions only lead to weak (<20 kcal mol^{-1} $Å^{-2}$) variations in the force constant of the residues, stressing the protein along the 75/171 direction results in a strong rigidity decrease of Ser^{27} (−57 kcal mol^{-1} $Å^{-2}$), Lys^{34} (−31 kcal mol^{-1} $Å^{-2}$), Glu^{119} (−64 kcal mol^{-1} $Å^{-2}$), and Val^{120} (-78 kcal mol^{-1} $Å^{-2}$). The initial distances between the C_α atoms from the 75/171, 40/171, and 40/130 residue pairs are 35.7 Å, 27 Å, and 28 Å, respectively. This indicates that the stress exerted by the constraint spring, which has an equilibrium length of 150 Å, should intrinsically be less pronounced for the 75/171 direction than for the 40/171 and 40/130 directions of load. However, the highly anisotropic architecture of the protein leads to the opposite effect, with the 75/171 direction inducing the most important changes in the enzyme's mechanics.

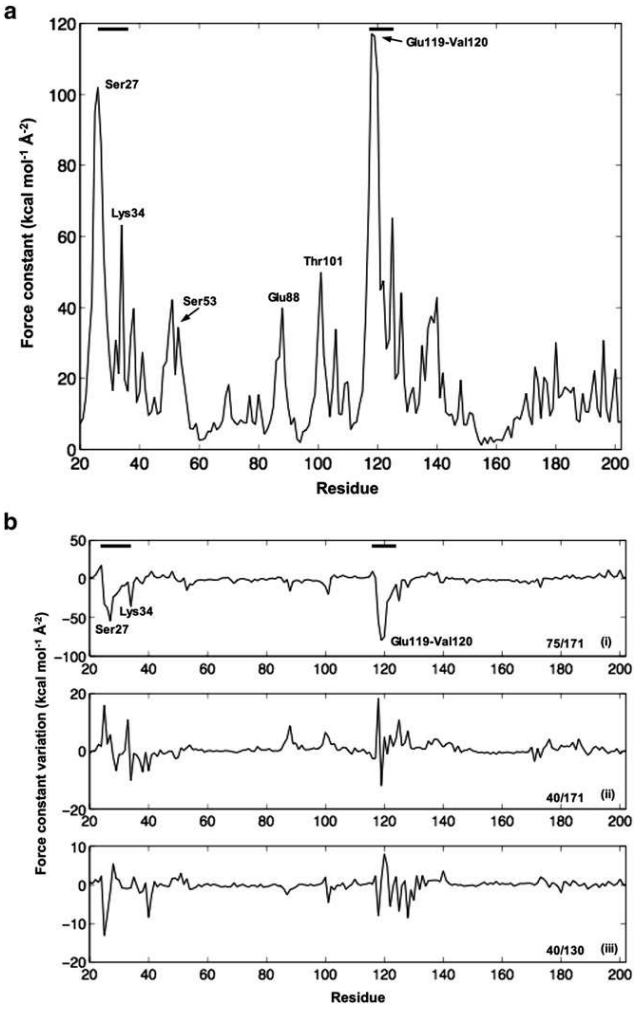


FIGURE 2 (*a*) Rigidity profile of GK when no mechanical stress is applied on the protein. (*b*) Variation of the force-constant profile upon applying an external constraint on the protein in pulling directions 75/171 (*upper*), 40/171 (*middle*), and 40/130 (*lower*). The black horizontal bars at the upper left of Fig. 2, *a* and *b*,indicate the position of the P-loop and the β7-sheet along the sequence.

Dynamics of the constrained protein

In a second step, we used the PCA approach to compute the inner product of the 10 first eigenvectors of a constrained trajectory with the eigenvectors of the relaxed trajectory. The resulting matrices for the 75/171, 40/171, and 40/130 pulling directions are plotted in Fig. 3, *a–c*, respectively.

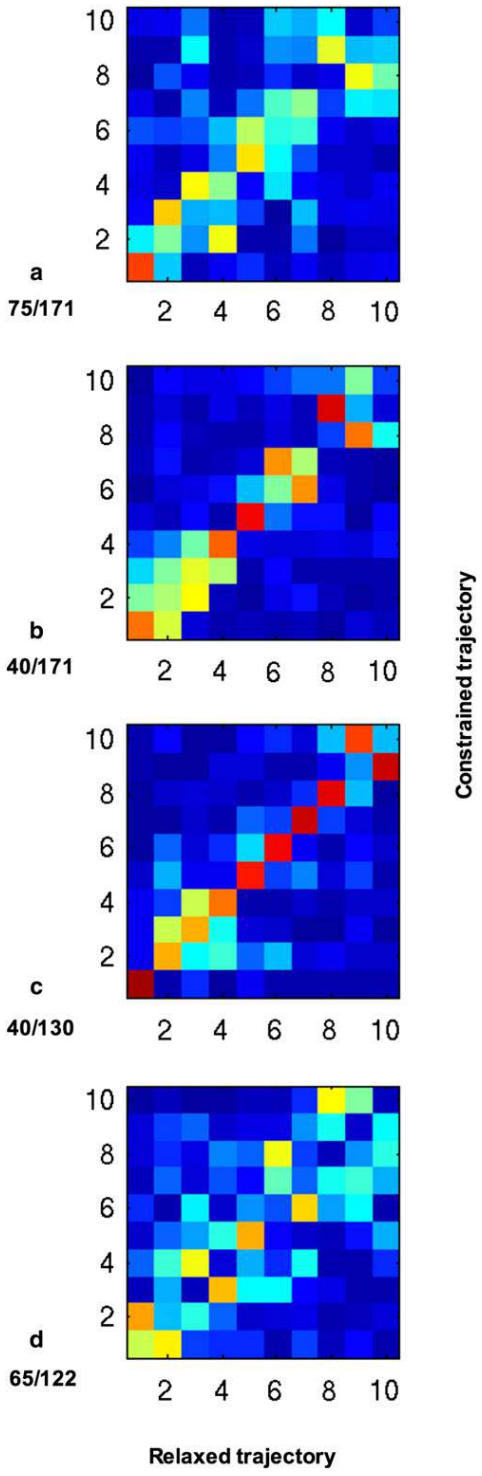


FIGURE 3 Inner-product matrices of the 10 first eigenvectors of the constrained trajectories over the relaxed trajectory. The pulling directions are (*a*) 75/171, (*b*) 40/171, (*c*) 40/130, and (*d*) 65/122. The color scale ranges from 0 to 1.

Table 1 summarizes the overlap between the covariance matrices (CMOs) of the relaxed and constrained trajectories. We can see how the application of an external stress along the 40/130 direction feebly modifies the modes of motion of the protein (CMO of 0.83), thus resulting in an almost diagonal matrix. On the other hand, applying an external constraint along the 75/171 and 40/171 directions induces some important disruptions of the protein's dynamics, but in different ways. Although pulling the protein along the 75/171 direction leads to a generally more important perturbation of GK movements compared with the 40/171 direction, with CMOs of 0.65 and 0.80, respectively, it turns out that the main functional mode of motion of the enzyme, corresponding to its opening and closing around the GMP and ATP binding sites, is more preserved for the 75/171 than for the 40/171 direction. This is shown by the projections of the first eigenvector of the constrained trajectory on the first eigenvector of the relaxed trajectory, which amount to overlaps of 0.83 and 0.77, respectively. This variation in the disruption of the enzyme dynamics is easily understandable, since the 75/171 direction actually coincides with opening and closing motions of GK, whereas pulling the protein via a constraint spring anchored on Cys^{40} and Arg^{171} introduces a new direction of motion with a component orthogonal to the main enzymatic movement.

TABLE 1 Overlap of constrained trajectories with relaxed trajectory along the 10 first modes of motions of the protein.

Pulling direction	Covariance matrix overlap (CMO)	Overlap of first eigenvectors
Thr^{75}-Arg^{171}*	0.65	0.83
Cys^{40}-Arg^{171}*	0.80	0.77
Cys^{40}-Lys^{130}*	0.83	0.97
Asp^{65}-Leu^{122}	0.66	0.55

*Asterisks identify the pulling directions experimentally tested by Tseng et al. (48)

Prediction of hotspots on the protein surface

We performed a systematic search of the protein mechanical response to the application of an external constraint by probing 236 new nonredundant pulling directions via residues anchored all over the surface of the protein. The resulting variations of the force constant of each residue are plotted in Fig. 4. From the qualitative point of view, most directions of load result in variations of the rigidity profile that are similar to those previously observed for the experimentally tested directions. Once again, the most important changes in the force constants of the residues occur in the P-loop and β7-sheet areas, which usually undergo a strong increase in flexibility.

The pulling direction that led to the most important perturbation of the rigidity profile (in terms of both the maximum force constant variation and the average perturbation of the profile) was formed by Asp^{65} and Leu^{122} (see Fig. 5). As can be seen in Fig. 5 (*lower*), the application of a constraint spring between these two residues induces

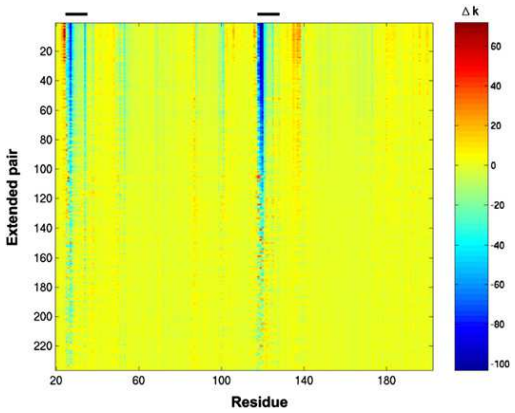


FIGURE 4 Distribution of the force-constant variation over all the pulling directions that have been modeled. Δk is expressed in kcal mol^{-1} Å^{-2}. The black horizontal bars above the figure indicate the position of the P-loop and the β7-sheet along the sequence. The extended pairs have been ordered starting with the direction leading to the largest average perturbation, $\langle \mathrm{abs}(\Delta k) \rangle$, of the GK rigidity profile. A negative/positive, value of Δk denotes a decrease/increase, of the rigidity of the residue.

an important disruption of the GMP binding site, with force-constant decreases beyond 100 kcal mol^{-1} Å^{-2} for Glu^{119} and Val^{120}, which is not surprising since the anchor residues actually surround the catalytic site. It is of interest that this

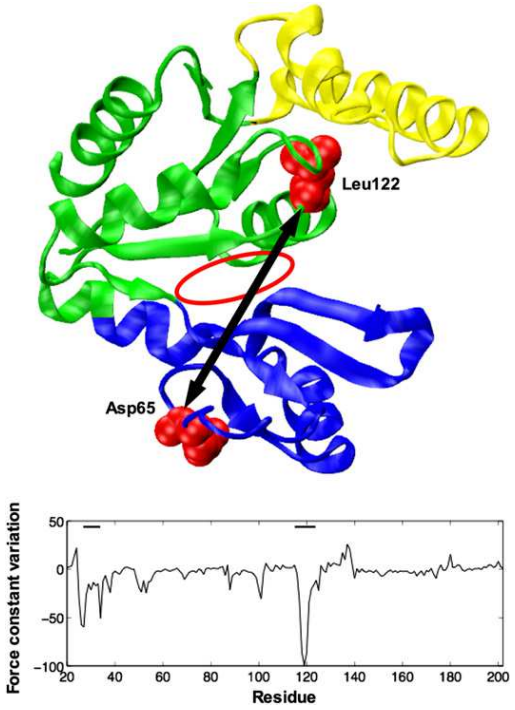


FIGURE 5 (*Upper*) Cartoon representation of GK with the 65/122 pulling direction. The ellipse indicates the location of the GMP-binding site. (Lower) Variations in force constant (kcal mol^{-1} Å^{-2}) as the protein moves along the 65/122 pulling direction. The black horizontal bars above the figure indicate the position of the P-loop and the β7-sheet along the sequence.

new pulling direction also strongly disturbs the enzyme's dynamics. As we can see from Fig. 3 *d* and Table 1, the 65/122 direction yields a CMO between the constrained and the relaxed trajectories of 0.66, whereas the projection of the first eigenvector is now reduced to 0.55, a much lower value than previously obtained for the experimental directions of load.

DISCUSSION AND CONCLUSION

In the ASP experiment, protein-DNA chimeras are used to strain the conformation of a protein, thus potentially providing further insight into the mechanism of allosteric control of biological function. In their work, Tseng et al. (48) applied a mechanical constraint at three different locations on the surface of GK, Thr^{75}/Arg^{171}, Cys^{40}/Arg^{171}, and Cys^{40}/Lys^{130}. These experiments yielded different results in terms of changes of the enzymatic activity: The 75/171 pulling direction induced a decrease in binding affinity for GMP, thus increasing the Michaelis-Menten constant, K_G, which was measured by GMP titration experiments, but having little or no effect on K_A (the binding affinity for ATP) and k_{cat} (the catalytic rate of the enzymatic reaction). For the 40/171 direction, Tseng et al. (48) observed a decrease of k_{cat}, whereas here, K_A and K_G were not affected. Finally, for the 40/130 pulling direction, no noticeable effect was observed for K_G, K_A, or k_{cat}. In this study, we combined a coarse-grained protein representation and BD simulations to investigate the mechanical and dynamical response of GK when an external stress is applied on the protein. During the simulations, the spring network oscillates around its equilibrium state within a limited range, with the deformations amounting to an ~1-Å root-mean-square deviation from the average conformation of the protein. Our model is therefore well adapted to describe the experiment of Tseng et al., where the protein's structure undergoes very few changes.

The force-constant profiles, which were obtained for trajectories with and without the application of a mechanical constraint, present rigidity peaks that correspond to residues belonging to the ligand-binding sites, thus stressing once more the importance of the catalytic site's stiffness for enzymatic activity (41,42). From the qualitative point of view, the force-constant variations observed for the protein under stress were mainly located around the P-loop and the β7-sheet. Quantitatively, however, only the 75/171 pulling direction led to an important decrease of the rigidity of residues Ile^{118} to Asp^{121}, which belong to the GMP binding site and do normally form interactions with the ligand in the closed conformation of the protein (67). Since such catalytic residues require an enhanced rigidity for the execution of their biological function (68–70), this disruption of the mechanical properties of the GMP binding site provides a first explanation for the decrease of the GMP binding affinity observed experimentally with the 75/171 chimera.

We then used PCA to study the variation in protein dynamics induced by the external stress. The resulting inner-product matrices obtained for the first 10 eigenvectors indicate that the 75/171 pulling direction leads to the most important perturbations of the general enzyme dynamics. However, if we focus only on the first mode of motion of the protein, which corresponds to the opening and closing movement of the LID and GMP domains over the CORE domain, and which is essential for GK to perform its catalytic function, it turns out that this mode is most perturbed when load is applied along the 40/171 direction. Once again this result is in agreement with the experimental data, where the 40/171 mutant alone presented a decrease of its catalytic rate, k_{cat}. All in all, the variations in enzymatic activity observed via the ASP experiments can either be related to some local mechanical perturbation of a substrate binding site (in the case of the GMP binding affinity), or to more global changes in the protein large-amplitude movements (for the catalytic rate constant).

Eventually, since Tseng et al. raised the question of the prediction of hotspots at the surface of the protein where a mechanical perturbation would produce a large functional response, we scanned 236 nonredundant locations on the protein surface. From a mechanical point of view, all pairs led to similar changes in the protein properties, with a rigidity decrease in the P-loop and β7-sheet regions. We were yet able to single out a specific residue pair that would be interesting to study experimentally, Asp^{65}-Leu^{122}. Since these two residues surround the GMP binding site, pulling in the 65/122 direction should yield an important disruption of its rigidity. It is also noteworthy that the neighboring Asp^{121} residue is implicated in GMP binding and initiation of the enzyme's closure (67). From a dynamical point of view, it appears that this direction of load also induces a strong perturbation of the protein's first mode of motion. This means that, were experiments performed on a 65/122 mutant, one should observe a decrease in both the enzyme's binding affinity for GMP and its catalytic rate. We furthermore tried to assess from the results of the systematic scan whether a particular pulling direction could specifically disrupt the ATP binding site while leaving the GMP binding site intact. This would result in a protein-DNA chimera where K_A, but not K_G, would be affected. In terms of mechanics, this meant finding a direction of load leading to a decrease in rigidity of the protein around the P-loop but not in the β7-sheet area. However, we could not find any residue pairs that would satisfy such a criterion, and it seems that the mechanical properties of these two elements of GK are tightly coupled and cannot be modulated independent of each other. To mechanically separate these subunits, one would probably have to disrupt the set of interactions that bind together the α1-helix and the β1- and β7-sheets via site-directed mutagenesis. In this perspective, residues Leu^{26}, Lys^{34}, Val^{38}, Leu^{117}, and Glu^{119}, whose side chains are directed toward the center of the CORE domain, appear to be potential candidates. It is interesting that all these residues also present important force constants (>35 kcal mol^{-1} Å^{-2}) in the rigidity profile of GK, which supports the idea that they might play an important part in the enzyme's structural stability. However, since Lys^{34} and Glu^{119} are also involved in ATP- and GMP-binding, respectively, it is unlikely that one could perturb the protein's intrinsic mechanics without disturbing its biological activity as well.

Altogether, we showed how a simple protein representation combined with BD simulations can yield a molecular level picture of the way the application of an external constraint on the GK enzyme perturbs its mechanical properties and dynamics. These perturbations can then be related to the changes observed experimentally in parameters K_G, K_A, and k_{cat} of the enzymatic reaction, helping us to understand the origin of the anisotropic functional response of the protein to various pulling directions. It is interesting to note that the apparent origin of the variations in thermodynamic parameter K_G is a disturbance of the protein's mechanical properties around its GMP binding site, whereas the changes in kinetic parameter k_{cat} arise from a perturbation of the protein dynamics, more precisely from the disruption of its first mode of motion, which defines the opening and closing movement performed by the protein when it undergoes a catalytic cycle. Our model can also be used in a predictive outlook. From a general search on the protein surface, we could suggest a new direction of load that should lead to the simultaneous perturbation of the two parameters K_G and k_{cat}, an effect not observed with the protein-DNA chimeras produced thus far.

SUPPORTING MATERIAL

Three figures and one table are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(10)01173-2.

This work was funded by the French Agency for Research (grants ANR-06-PCVI-0025 and ANR-07-CIS7-003).

REFERENCES

1. Daniel, R. M., R. V. Dunn, …, J. C. Smith. 2003. The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.* 32:69–92.
2. Parak, F. G. 2003. Physical aspects of protein dynamics. *Rep. Prog. Phys.* 66:103–129.
3. Bustamante, C., Y. R. Chemla, …, D. Izhaky. 2004. Mechanical processes in biochemistry. *Annu. Rev. Biochem.* 73:705–748.
4. Hammes-Schiffer, S., and S. J. Benkovic. 2006. Relating protein motion to catalysis. *Annu. Rev. Biochem.* 75:519–541.
5. Charvin, G., J. F. Allemand, …, V. Croquette. 2004. Twisting DNA: single molecule studies. *Contemp. Phys.* 45:383–403.
6. Koster, D. A., V. Croquette, …, N. H. Dekker. 2005. Friction and torque govern the relaxation of DNA supercoils by eukaryotic topoisomerase IB. *Nature.* 434:671–674.
7. Moffitt, J. R., Y. R. Chemla, …, C. Bustamante. 2008. Recent advances in optical tweezers. *Annu. Rev. Biochem.* 77:205–228.

8. Brockwell, D. J. 2007. Probing the mechanical stability of proteins using the atomic force microscope. *Biochem. Soc. Trans.* 35:1564–1568.
9. Puchner, E. M., and H. E. Gaub. 2009. Force and function: probing proteins with AFM-based force spectroscopy. *Curr. Opin. Struct. Biol.* 19:605–614.
10. Galera-Prat, A., A. Gómez-Sicilia, …, M. Carrión-Vázquez. 2010. Understanding biology by stretching proteins: recent progress. *Curr. Opin. Struct. Biol.* 20:63–69.
11. Kumar, S., and M. S. Li. 2010. Biomolecules under mechanical force. *Phys. Rep.* 486:1–74.
12. Brockwell, D. J., E. Paci, …, S. E. Radford. 2003. Pulling geometry defines the mechanical resistance of a β-sheet protein. *Nat. Struct. Biol.* 10:731–737.
13. Carrion-Vazquez, M., H. Li, …, J. M. Fernandez. 2003. The mechanical stability of ubiquitin is linkage dependent. *Nat. Struct. Biol.* 10:738–743.
14. Dietz, H., F. Berkemeier, …, M. Rief. 2006. Anisotropic deformation response of single protein molecules. *Proc. Natl. Acad. Sci. USA.* 103:12724–12728.
15. Nome, R. A., J. M. Zhao, …, N. F. Scherer. 2007. Axis-dependent anisotropy in protein unfolding from integrated nonequilibrium single-molecule experiments, analysis, and simulation. *Proc. Natl. Acad. Sci. USA.* 104:20799–20804.
16. Puchner, E. M., A. Alexandrovich, …, M. Gautel. 2008. Mechanoenzymatics of titin kinase. *Proc. Natl. Acad. Sci. USA.* 105:13385–13390.
17. Oberhauser, A. F., and M. Carrión-Vázquez. 2008. Mechanical biochemistry of proteins one molecule at a time. *J. Biol. Chem.* 283:6617–6621.
18. Monod, J., J. P. Changeux, and F. Jacob. 1963. Allosteric proteins and cellular control systems. *J. Mol. Biol.* 6:306–329.
19. Goodey, N. M., and S. J. Benkovic. 2008. Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.* 4:474–482.
20. Alberts, B., D. Bray, …, J. Watson. 1994. Molecular Biology of the Cell. Garland Science, New York.
21. Choi, B., G. Zocchi, …, L. J. Perry. 2005. Artificial allosteric control of maltose binding protein. *Phys. Rev. Lett.* 94:038103.
22. Choi, B., G. Zocchi, …, L. Jeanne Perry. 2005. Allosteric control through mechanical tension. *Phys. Rev. Lett.* 95:078102.
23. Zocchi, G. 2009. Controlling Proteins Through Molecular Springs. *Annu. Rev. Biophys.* 38:75–88.
24. Choi, B., and G. Zocchi. 2007. Guanylate kinase, induced fit, and the allosteric spring probe. *Biophys. J.* 92:1651–1658.
25. Paci, E., and M. Karplus. 2000. Unfolding proteins by external forces and temperature: the importance of topology and energetics. *Proc. Natl. Acad. Sci. USA.* 97:6521–6526.
26. Lu, H., and K. Schulten. 2000. The key event in force-induced unfolding of Titin's immunoglobulin domains. *Biophys. J.* 79:51–65.
27. Gräter, F., J. H. Shen, …, H. Grubmüller. 2005. Mechanically induced titin kinase activation studied by force-probe molecular dynamics simulations. *Biophys. J.* 88:790–804.
28. Ortiz, V., S. O. Nielsen, …, D. E. Discher. 2005. Unfolding a linker between helical repeats. *J. Mol. Biol.* 349:638–647.
29. Lee, E. H., J. Hsin, …, K. Schulten. 2009. Discovery through the computational microscope. *Structure.* 17:1295–1306.
30. Nielsen, S. O., C. F. Lopez, …, M. L. Klein. 2004. Coarse grain models and the computer simulation of soft materials. *J. Phys. Cond. Mat.* 16:R481–R512.
31. Tozzini, V., J. Trylska, …, J. A. McCammon. 2007. Flap opening dynamics in HIV-1 protease explored with a coarse-grained model. *J. Struct. Biol.* 157:606–615.
32. Monticelli, L., S. K. Kandasamy, …, S. J. Marrink. 2008. The MARTINI coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.* 4:819–834.
33. Noid, W. G., J. W. Chu, …, H. C. Andersen. 2008. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* 128:244114.
34. Noid, W. G., P. Liu, …, G. A. Voth. 2008. The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models. *J. Chem. Phys.* 128:244115.
35. Tozzini, V. 2010. Multiscale modeling of proteins. *Acc. Chem. Res.* 43:220–230.
36. Tirion, M. M. 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* 77:1905–1908.
37. Bahar, I., A. R. Atilgan, and B. Erman. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* 2:173–181.
38. Hinsen, K., A. Thomas, and M. J. Field. 1999. Analysis of domain motions in large proteins. *Proteins.* 34:369–382.
39. Atilgan, A. R., S. R. Durell, …, I. Bahar. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80:505–515.
40. Tama, F., W. Wriggers, and C. L. Brooks, 3rd. 2002. Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *J. Mol. Biol.* 321:297–305.
41. Sacquin-Mora, S., and R. Lavery. 2006. Investigating the local flexibility of functional residues in hemoproteins. *Biophys. J.* 90: 2706–2717.
42. Sacquin-Mora, S., E. Laforet, and R. Lavery. 2007. Locating the active sites of enzymes using mechanical properties. *Proteins.* 67:350–359.
43. Yang, L., G. Song, and R. L. Jernigan. 2009. Protein elastic network models and the ranges of cooperativity. *Proc. Natl. Acad. Sci. USA.* 106:12347–12352.
44. Yang, L., G. Song, …, R. L. Jernigan. 2008. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure.* 16:321–330.
45. Dietz, H., and M. Rief. 2008. Elastic bond network model for protein unfolding mechanics. *Phys. Rev. Lett.* 100:098101.
46. Eyal, E., and I. Bahar. 2008. Toward a molecular understanding of the anisotropic response of proteins to external forces: insights from elastic network models. *Biophys. J.* 94:3424–3435.
47. Sacquin-Mora, S., and R. Lavery. 2009. Modeling the mechanical response of proteins to anisotropic deformation. *ChemPhysChem.* 10:115–118.
48. Tseng, C. Y., A. Wang, and G. Zocchi. 2010. Mechano-chemistry of the enzyme Guanylate Kinase. *Eur. Phys. Lett.* 91:18005.
49. Oeschger, M. P., and M. J. Bessman. 1966. Purification and properties of guanylate kinase from *Escherichia coli*. *J. Biol. Chem.* 241: 5452–5460.
50. Blaszczyk, J., Y. Li, …, X. Ji. 2001. Crystal structure of unligated guanylate kinase from yeast reveals GMP-induced conformational changes. *J. Mol. Biol.* 307:247–257.
51. Sekulic, N., L. Shuvalova, …, A. Lavie. 2002. Structural characterization of the closed conformation of mouse guanylate kinase. *J. Biol. Chem.* 277:30236–30243.
52. Tozzini, V. 2005. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* 15:144–150.
53. Zacharias, M. 2003. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.* 12:1271–1282.
54. Tseng, C. Y., A. Wang, …, A. J. Levine. 2009. Elastic energy of protein-DNA chimeras. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 80:061912.
55. García, A. E. 1992. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* 68:2696–2699.
56. Amadei, A., A. B. M. Linssen, and H. J. C. Berendsen. 1993. Essential dynamics of proteins. *Proteins.* 17:412–425.

57. Case, D. A. 1994. Normal-mode analysis of protein dynamics. *Curr. Opin. Struct. Biol.* 4:285–290.

58. Amadei, A., A. B. M. Linssen, …, H. J. Berendsen. 1996. An efficient method for sampling the essential subspace of proteins. *J. Biomol. Struct. Dyn.* 13:615–625.

59. Berendsen, H. J. C., D. Vanderspoel, and R. Vandrunen. 1995. GROMACS: a message-passing parallel molecular-dynamics implementation. *Comput. Phys. Commun.* 91:43–56.

60. Lindahl, E., B. Hess, and D. van der Spoel. 2001. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* 7:306–317.

61. Van Der Spoel, D., E. Lindahl, …, H. J. Berendsen. 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26:1701–1718.

62. Miller, S., J. Janin, …, C. Chothia. 1987. Interior and surface of monomeric proteins. *J. Mol. Biol.* 196:641–656.

63. Hubbard, S. J. 1992. ACCESS: A Program for Calculating Accessibilities. Department of Biochemistry and Molecular Biology, University College of London, London.

64. Dreusicke, D., and G. E. Schulz. 1986. The glycine-rich loop of adenylate kinase forms a giant anion hole. *FEBS Lett.* 208:301–304.

65. Leipe, D. D., E. V. Koonin, and L. Aravind. 2003. Evolution and classification of P-loop kinases and related proteins. *J. Mol. Biol.* 333: 781–815.

66. Stehle, T., and G. E. Schulz. 1992. Refined structure of the complex between guanylate kinase and its substrate GMP at 2.0 Å resolution. *J. Mol. Biol.* 224:1127–1141.

67. Hible, G., P. Christova, …, J. Cherfils. 2006. Unique GMP-binding site in *Mycobacterium tuberculosis* guanosine monophosphate kinase. *Proteins.* 62:489–500.

68. Bartlett, G. J., C. T. Porter, …, J. M. Thornton. 2002. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* 324:105–121.

69. Yuan, Z., J. Zhao, and Z. X. Wang. 2003. Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng.* 16: 109–114.

70. Yang, L. W., and I. Bahar. 2005. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure.* 13:893–904.

71. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27–38.

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

ARTICLE

pubs.acs.org/JACS

Frontier Residues Lining Globin Internal Cavities Present Specific Mechanical Properties

Anthony Bocahut,[†] Sophie Bernad,[‡] Pierre Sebban,[‡,§] and Sophie Sacquin-Mora*[,†]

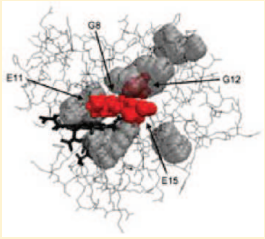
[†]Laboratoire de Biochimie Théorique, UMR 9080 CNRS, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005 Paris, France

[‡]Laboratoire de Chimie Physique, CNRS UMR8000, Bât. 350, Université Paris-sud, 91405 Orsay, France

[§]Université des Sciences et des Technologies de Hanoi, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam

Supporting Information

ABSTRACT: The internal cavity matrix of globins plays a key role in their biological function. Previous studies have already highlighted the plasticity of this inner network, which can fluctuate with the proteins breathing motion, and the importance of a few key residues for the regulation of ligand diffusion within the protein. In this Article, we combine all-atom molecular dynamics and coarse-grain Brownian dynamics to establish a complete mechanical landscape for six different globins chain (myoglobin, neuroglobin, cytoglobin, truncated hemoglobin, and chains α and β of hemoglobin). We show that the rigidity profiles of these proteins can fluctuate along time, and how a limited set of residues present specific mechanical properties that are related to their position at the frontier between internal cavities. Eventually, we postulate the existence of conserved positions within the globin fold, which form a mechanical nucleus located at the center of the cavity network, and whose constituent residues are essential for controlling ligand migration in globins.



INTRODUCTION

The globin superfamily is found in all kingdoms of life, and its members can perform a large variety of functions such as NO scavenging, enzymatic activities, oxygen sensing, and, of course, O_2 transport and storage.[1−5] Interestingly, their sequence can be extremely variable, with globins presenting less than 10% homology,[6] and they are best characterized by their common structural feature. This typical 3D fold of a small number of α-helices, named the globin fold, protects a noncovalently bound heme group and alows reversible ligand binding. Despite over 50 years of intensive research,[7] globins still represent a fascinating subject, their structural and functional properties being far from fully understood,[8,9] and with recently discovered members, such as neuroglobin or cytoglobin, whose physiological function has remained elusive until today.[5,10,11]

The internal cavity network located in the matrix of globular proteins usually plays a key role in ligand migration and for the control of protein function.[12−16] In the case of globins, the diffusion pathways of various small ligands have been extensively studied for over 30 years,[17−28] showing great variability among the different members of the family.[29] In their work,[29] Cohen and Schulten also noted that, despite this multiplicity of ligand migrations pathways that could be observed among globins, some specific positions within the globin fold could actually present a propensity to be located near a ligand passageway. In a previous study on human neuroglobin (Ngb),[30] we showed that the mechanical properties of the residues lying at the border between two internal cavities could be related to the ligand migration pathways that were observed via metadynamics simulations. In a recently published paper on myoglobin (Mgb),[31] Scorciapino et al. identified a set of key residues likely to work as switches regulating ligand migration from one cavity to the other. After noting that these "frontier" residues did occupy similar positions along both the Mgb and Ngb sequences, we sat about investigating their mechanical properties in an extended set of globin chains comprising also cytoglobin (Cgb), truncated hemoglobin (Tr. Hb), and the α and β chains of human hemoglobin (Hb). Although the common general features of globins dynamics have already been studied,[32,33] we chose here to focus on frontier residues, to understand how their mechanical properties can affect ligand migration within the protein cavity network. In this perspective, we use an approach combining all-atom classical molecular dynamics (MD) and coarse-grain Brownian dynamics simulations to draw a complete picture of globins mechanics and show how a limited set of residues might be playing a key role for ligand diffusion in the protein matrix.

MATERIALS AND METHODS

The starting coordinates employed for the simulations were taken from the experimental X-ray structure of each globin. In the case of human Ngb, we took the B chain of the 1OJ6[34] PDB file (with 1.95 Å resolution), and we performed three mutations in silico (G46C, S55C, and S120C) to retrieve the wild-type cysteines, which are not present in the crystal. For the other globins, we chose the following PDB entries:

Received: March 22, 2011
Published: May 09, 2011

dx.doi.org/10.1021/ja202587a | J. Am. Chem. Soc. 2011, 133, 8753–8761

Table 1. Clustering of the 17 501 Conformers Obtained for Each of the 35 ns MD Simulations of Six Representative Globins

Human Neuroglobin (10J6, B Chain), 151 Residues				
NGB0	NGB1	NGB2	NGB3	NGB4
20%	15%	4%	16%	44%
Myoglobin (1YMB), 153 Residues				
MGB0	MGB1	MGB2	MGB3	MGB4
53%	18%	17%	11%	2%
Truncated Hemoglobin (1IDR, B Chain), 126 Residues				
THB0	THB1	THB2	THB3	THB4
45%	17%	16%	15%	7%
Cytoglobin (2DC3, A Chain), 155 Residues				
CGB0	CGB1	CGB2	CGB3	CGB4
16%	one conformer	44%	13%	27%
αHemoglobin (2HHB, A Chain), 141 Residues				
AHB0	AHB1	AHB2	AHB3	AHB4
31%	14%	16%	30%	9%
βHemoglobin (2HHB, B Chain), 146 Residues				
BHB0	BHB1	BHB2	BHB3	BHB4
51%	12%	37%	one conformer	one conformer

Horse heart Mgb from 1YMB[35] at 2.8 Å resolution; Human Cgb from 2DC3[36] at 1.68 Å resolution (A chain); Tr. Hb of *Mycobacterium tuberculosis* from 1IDR[18] at 1.9 Å resolution (B chain); and human Hb from 2HHB[37] at 1.74 Å resolution (chains A for αHb and B for βHb).

Classical Molecular Dynamics. MD simulations were performed with the Gromacs[38−40] software package using the OPLS all atoms force field.[41] Quantum chemical calculations with Gaussian[42] were performed to determine the charges of the hexacoordinated heme group (Ngb, Cytg, TrHb) and pentacoordinated heme group (Mgb, αHb, βHb) using B3LYP[43] and the 6-31G* basis set. The other force field parameters for the prosthetic group were taken from previous studies done on Mgb.[44] The protein was solvated in a cubic box of side length 78 Å, using periodic boundary conditions, with explicit single-point charge water molecules.[45] When necessary, Na^+ ions (from two to six) were added to neutralize the system, which contained between 47 000 and 52 000 atoms depending on the globin under study. All simulations were performed at 1 atm and 300 K, maintained with the Berendsen barostat and thermostat.[46] Long-range electrostatic interactions were treated using the Particle Mesh Ewald (PME) method,[47] with a grid spacing of 0.12 nm and a nonbond pair list cutoff of 9.0 Å with an updating of the pair list every five steps. We could choose a time step of 2 fs by constraining bond lengths involving H atoms with the LINCS algorithm.[48] The solvent was first relaxed by an energy minimization, which was followed by a 100 ps equilibration step under restraint, and then heated slowly until 300 K; 50 ns production runs were eventually performed from which the last 35 ns were kept for analysis. The g_cluster algorithm from the Gromacs suite was then used to obtain five representative structures for each globin over the simulation production period (see Table 1). We used the single linkage method, where a new structure is added to the cluster when the distance between two conformations is less than a chosen cutoff, and employed a different clustering cutoff for each globin, depending on the weight of the system. For Mgb we used a 0.0788 cutoff; 0.078 nm for Ngb; 0.0763 nm for Cgb; 0.0802 nm for TrHb; 0.0777 nm for αHb; and 0.0795 nm for βHb. The 30 resulting structures (listed in Table 1) are identified via a three-letter code indicating the original globin (MGB, NGB, CGB, THB, AHB, or BHB) and a number (from 0 to 4) for the cluster. MGB0, for example, corresponds to the first clusterized structure obtained for horse heart Mgb.

Finally, the online software Pocket-Finder (http://www.modelling.leeds.ac.uk/pocketfinder/)[49] was used for detecting cavities in the various globin structures that were produced and calculating their volumes. These calculations were performed on the clusterized structures with their prosthetic group but in the absence of ligand.

Brownian Dynamics Simulations. BD simulations have been carried out on the globins clusterized structures using the ProPHet (Probing Protein Heterogeneity) program.[50−52] The simulations used a coarse-grained protein model, in which each amino acid is represented by one pseudoatom located at the Cα position, and either one or two (for larger residues) pseudoatoms replacing the side chain (with the exception of Gly).[53] Interactions between the pseudoatoms are treated according to the standard elastic network model;[54] that is, all pseudoatoms lying closer than 9 Å are joined with quadratic springs having the same force constant of 0.6 kcal mol^{-1} $Å^{-2}$. Springs are assumed to be relaxed in the reference conformation of the protein, derived either from the crystallographic data or from the clusterized structures produced by the MD simulations. Following earlier studies, which showed how ligands as large as a heme group actually had little influence on calculated force constants,[50,51] we chose not to include the prosthetic group in the protein representation. The simulations use an implicit solvent representation via the diffusion and random displacement terms in the equation of motion,[55] and hydrodynamic interactions are included through the diffusion tensor.[56]

From the positional fluctuations resulting from BD simulations, carried out for 100 000 steps at a temperature of 300 K, effective force constants for displacing each particle i are calculated as

$$k_i = \frac{3k_B T}{\langle (d_i - \langle d_i \rangle)^2 \rangle} \quad (1)$$

where the brackets indicate an average taken over the whole simulation, k_B is the Boltzmann constant, and d_i is the average distance of particle i from the other particles j in the protein, excluding the pseudoatoms, which belong to the same residue m to which particle i belongs. Also, the distances between the Cα pseudoatom of residue m and the Cα pseudoatoms of the adjacent residues $m + 1$ and $m - 1$ are not included in the average. The force constant associated with each residue m is taken to be the average of the force constants calculated according to eq 1 for each of the pseudoatoms i forming this residue. Within this framework, the mechanical properties of the protein are described at the residue level by its "rigidity profile", that is, by the ordered sequence of the force constants calculated for each residue.

RESULTS

Globins Cavity Network. For the six studied globin chains, the clusterized structures were analyzed with the Pocket-Finder program, and the 10 main cavities detected in each of these structures are listed with their lining residues in Supporting Information Tables 1−6. Similarly to what we observed in our previous work on human neuroglobin,[30] the cavity network of each protein can show considerable reorganization from one cluster to the other, thus inducing large variations of the total volume of the cavities, which can range from 289 to 543 $Å^3$ for βHb and from 558 to 913 $Å^3$ for Mgb, if we take the two globins presenting the most extreme volume variations. A number of

these inner pockets can be related to the xenon cavities (Xe1, Xe2, Xe3, and Xe4) and the distal pocket (DP) that have been observed experimentally in sperm whale Mgb,[12] the phantom 1 cavity that was detected in the same protein by MD simulations,[21,22] the site 1 that was observed in a Xe adduct of human Hb,[57] or the numerous ligand exit pathways that could be identified via MD simulations on Mgb,[24−27] Hbs,[18,23] or Ngb;[30,34,58]

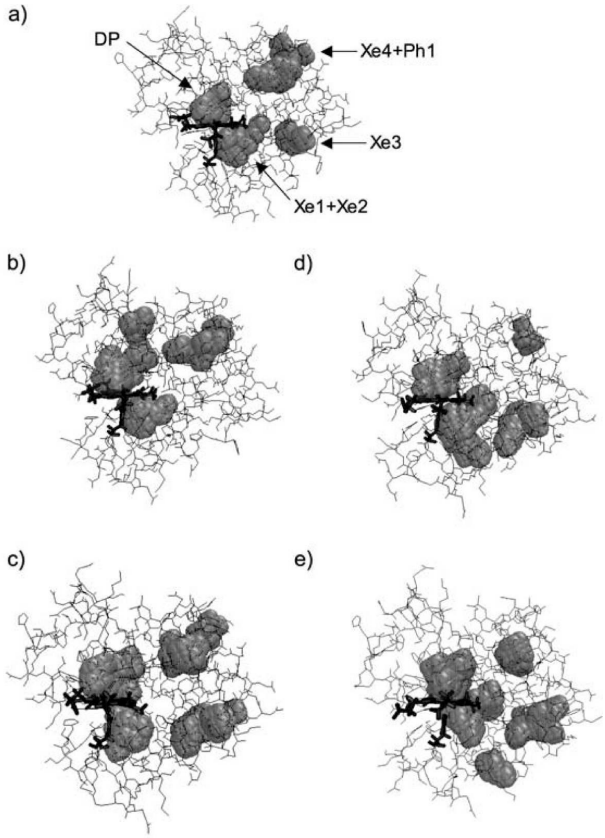


Figure 1. Representation of the five main cavities in the clusterized structures of horse-heart Mgb as detected by Pocket-Finder. (a) MGBO with arrows pointing to the standard Xe and Ph1 cavities, (b) MGB1, (c) MGB2, (d) MGB3, and (e) MGB4. This figure and Figures 5 and 6 were prepared using Visual Molecular Dynamics.[85]

see Figure 1 for a typical representation of the cavity network and its fluctuations in Mgb.

From the list of the cavity lining residues, we could define two subgroups of what we call frontier residues (FR), that is, residues lining two or more internal pockets in the protein. For a given globin, transient frontier residues (TFR) are located at the border between two cavities in only one of the five clusterized structures, while recurrent frontier residues (RFR) can be found in at least two of the clusterized structures; both groups are listed in tTable 2.

Globins Mechanical Properties. The force constant profiles obtained for the main structural cluster of each protein are plotted in Figure 2. Similar to what has been observed in our previous studies on hemoproteins[50] and neuroglobin,[30] the analogous aspect of the profiles reflects the α-helical globin fold, with α-helices appearing as grouped rigidity peaks along the protein sequence (see the shaded areas in Figure 2a) and flexible regions between, denoting in particular the CD and EF loops. In their work of 1999 made on 728 sequences of different globin subfamilies, Ptitsyn and Ting[59] identified 13 conserved heme-binding residues. It turns out 12 out of these 13 residues (which are indicated by empty squares in Figure 2a) actually correspond to local peaks in the proteins rigidity profiles. This suggests how important the tight binding of the prosthetic group is for the biological activity of the protein.[51] Likewise, the five residues forming the folding nucleus of globins (indicated by "●" in Figure 2a) correspond to rigidity peaks, thus underlying the strong correspondence between a protein mechanics and its functional and structural properties.

Even though the five clusterized structures obtained for each globin do not present important variations, with Cα rmsd's between two conformations that are always inferior to 2 Å and with an average value of 1.2 Å, these small structural changes are nonetheless sufficient to induce noticeable variations in the mechanical properties of a limited number of residues in the six globin chains. For every studied protein, we made a pairwise comparison of all five rigidity profiles, and for each residue we kept the maximum value that could be observed for its force constant variation. The resulting max(Δk) profiles are plotted in Figure 3. We then defined as "mechanically sensitive" (MS) those residues presenting a max(Δk) value over a given threshold of 10 kcal mol^{-1} Å^{-2} for Ngb, Mgb, Cgb, and αHb, 7 kcal mol^{-1} Å^{-2} for Tr. Hb, and 20 kcal mol^{-1} Å^{-2} for βHb. This procedure led to the selection of 8−14 residues for each globin that are listed in

Table 2. List of the Frontier Residues (Lining Two or More Internal Cavities), Which Were Obtained via Pocket-Finder for Each Globin[a]

Mgb	transient	10/14/28/29/39/42/46/61/66/78/81/82/86/89/93/97/100/101/11/115/118/123/142
	recurrent	17/21/25/43/64/65/69/72/75/76/77/86/99/104/107/138/146
Ngb	transient	41/42/71/82/99/102/103/105/111/147
	recurrent	27/28/38/68/72/75/85/89/92/95/96/101/106/109/110/113/133/136/137/140/144
Cgb	transient	30/41/42/45/49/84/88/93/135/143/156
	recurrent	31/34/56/60/81/85/86/92/102/106/109/124/127/128/131/134/151/154/157/158/161
Tr. Hb	transient	16/22/29/36/53/54/65/72/84/95/126
	recurrent	19/25/32/33/46/58/61/63/66/77/80/86/94/98/102/115/116/119/122
αHb	transient	21/25/30/58/95/102/117/121/130/132
	recurrent	14/17/24/29/33/43/48/55/62/63/66/101/105/106/109/117/125/129/133
βHb	transient	11/24/25/30/31/33/35/45/48/54/63/72/75/76/81/84/98/103/106/107/110/114/134/137/139/140
	recurrent	15/23/26/28/32/42/60/67/68/71/78/85/130

[a] These can be either transient (appearing in only one of the clusterized structures) or recurrent (present in two or more of the clusterized structures).

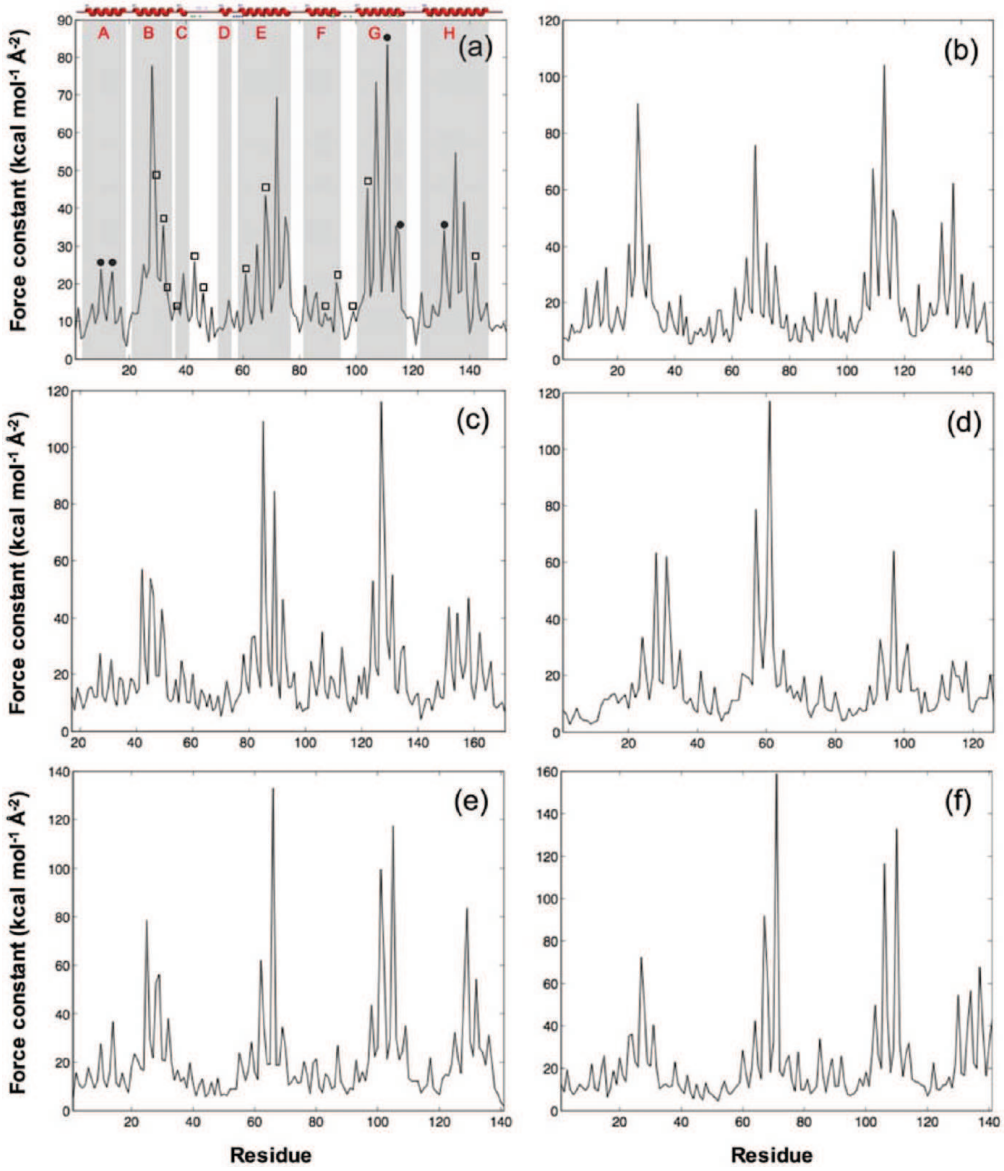


Figure 2. Rigidity profiles (in kcal mol^{-1} Å^{-2}) of the main cluster for the six globin chains under study. (a) Mgb, (b) Ngb, (c) Cgb, (d) Tr. Hb, (e) αHb, (f) βHb. In (a), the areas shaded in gray correspond to α-helices, as indicated by the red secondary structure plot at the top of the structure, the "□" indicate heme-binding conserved residues of globins (from left to right: Leu29-B10, Leu32-B13, Phe33-B14, Pro37-C2, Phe43-CD1, Phe46-CD4, Leu61-E4, Val68-E11, Leu89-F4, His93-F8, Ile99-FG5, Leu104-G5, and Ile142-H19), and the "●" indicate the conserved folding nucleus (from left to right: Val10-A8, Trp14-A12, Ile11-G12, Leu115-G16, and Met131-H8).

Table 3. Interestingly, these MS residues, which represent a subset of the rigid residues from the original rigidity profiles, systematically correspond to frontier residues in Mgb and Cgb. In the case of Ngb, Tr. Hb, αHb, and βHb, the few MS residues that are not frontier residues are nonetheless cavity lining residues, with the only exception of Ser112-G11 in Ngb.

Conservation of the Mechanical Properties along the Sequence. We used the clustalw[60] web server to align the sequences of the six globin chains under study. Despite the high mechanical similarity that could be observed in the rigidity profiles of Figure 2, theses sequences present relatively low identities, ranging from 17% to 50% (see Supporting Information Table 7). The multiple sequence alignment is presented in Figure 4 with the positions that are occupied by MS residues highlighted in red. We can see that most of these positions are indeed common to one or more globins, with the particular case of positions G8 and G12, which correspond to MS residues in all six chains. Two other positions that are extremely well conserved in terms of mechanical properties are E11, which presents a MS residue in all chains but Mgb, and E15, where Ngb is the only chain not showing a MS residue. However, the max(Δk) value of Val68-E11 in Mgb is actually right under the chosen cutoff with 9.14 kcal mol^{-1} Å^{-2}, which means that this residue does actually present mechanical sensitivity. In the case of Ile72-E15 of Ngb,

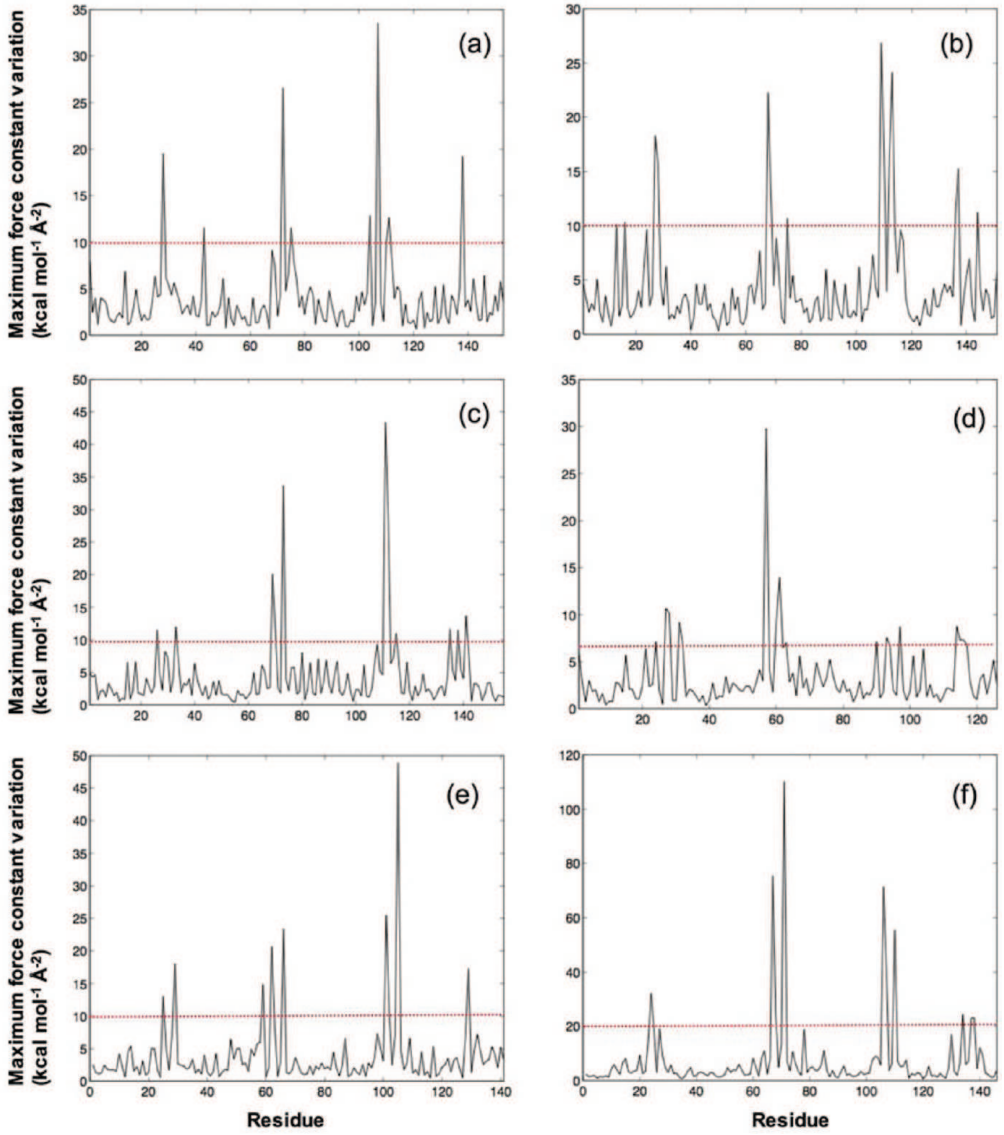


Figure 3. Maximum variation (in kcal mol^{-1} Å^{-2}) of the force constant upon changing the globin structure. (a) Mgb, (b) Ngb, (c) Cgb, (d) Tr. Hb, (e) αHb, (f) βHb. The red horizontal dotted line indicates the threshold value chosen for the selection of mechanically sensitive residues that are listed in Table 2.

Table 3. List of the Mechanically Sensitive Residues in Each of the Six Globins with Their Position along the Protein Fold[a]

Mgb	Val28-B9, Phe43-CD1, Leu72-E15, Ile75-E18, Leu104-G5, **Ile107-G8**, **Ile111-G12**, Phe138-H15
Ngb	Trp13-A12, Val16-A15, Leu27-B9, Phe28-B10, Val68-E11, Met69-E12, Ala75-E18, **Val109-G8**, Gly110-G9, Ser112-G11, **Leu113-G12**, Leu136-H11, Tyr137-H12, Met144-H19
Cgb	Gly42-B6, Phe49-B13, Val85-E11, Met86-E12, Leu89-E15, **Leu127-G8**, Ser128-G9, **Ile131-G12**, Trp151-H8, Leu154-H11
Tr. Hb	Ile25-B6, Val28-B9, Val29-B10, Phe32-B13, Gln58-E11, Phe61-E14, Phe62-E15, Ala64-E17, Phe91-G5, **Val94-G8**, **Leu98-G12**, Ile115-H8, Leu116-H9, Gly117-H10
αHb	Gly25-B6, Leu29-B10, Gly59-E8, Val62-E11, Leu66-E15, **Leu101-G8**, Ser102-G9, **Leu105-G12**, Leu129-H12
βHb	Gly24-B6, Val67-E11, Leu68-E12, Phe75-E15, **Leu106-G8**, Gly107-G9, **Leu110-G12**, Val134-H12, Val137-H15, Ala138-H16

[a] Positions common to more than one protein are underlined; positions common to all proteins are in bold.

this residue did show some specific mechanical properties in our previous work on human Ngb,[30] where we investigated its mechanical variations upon formation of an internal disulfide bond in the pentacoordinated state of the protein (which is the

Figure 4. Alignment of the six globin sequences. The first column in each block displays the PDB code and chain of the protein, and the last column shows the number of residues up to that line. Green annotations indicate the positions of the α-helices along the Mgb sequence, while mechanically sensitive residues are highlighted in red.

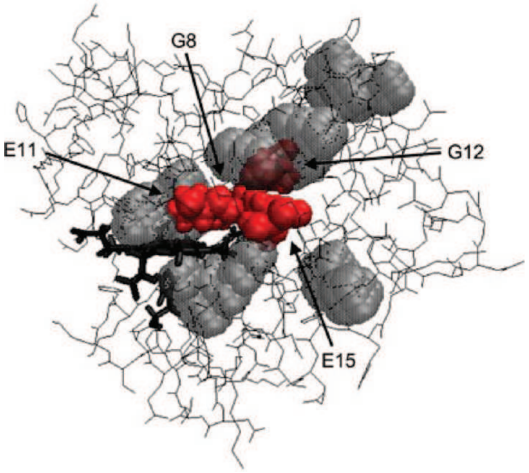


Figure 5. Conserved mechanical nucleus formed by positions E11, E15, G8, and G12 (in red) at the heart of the MGB0 structure.

ligand binding state), while the simulations in the current work were carried out on the hexacoordinated state of human Ngb. So eventually, we can select positions E11, E15, G8, and G12 along the globin sequence to define a conserved mechanical nucleus that appears to form a central gate system right at the heart of the globin fold and at the frontier of the DP, Xe2, and Xe4 cavities; see Figure 5.

DISCUSSION

The six globin chains in this study presented very similar rigidity profiles, thus reflecting the conservation of proteins dynamics within a structural family.[33,61,62] More interestingly, the variations of these profiles due to the proteins structural fluctuations are also comparable and allow us to select a restricted set of residues occupying that we called "mechanically sensitive" positions. A search in the literature shows that most of the positions bearing that label had already been highlighted as corresponding to cavity lining of frontier residues in numerous experimental or theoretical works on Mgb,[21,22,26,27,63] Ngb,[30,64−69] Cgb,[70−72] Tr. Hb,[18,73,74] and human Hb.[23] Interestingly, our results concur with data obtained from molecular dynamics performed using not only the OPLS force field like us,[27] but also Amber 95[31] and 99[73] or Charmm 22[23,29] and 27,[72] thus showing the robustness of molecular simulation studies for the investigation of protein properties.

The positions of the mechanical nucleus residues in particular have been shown to play an important role for ligand diffusion in various globins. For example, in the case of Mgb, several mutational studies focused on the importance of positions E11 and G8,[75−77] showing how the replacement of the isoleucine in G8 does not modify the protein's structure, but substantially affects ligand binding. In Scorciapino et al.'s work on Mgb breathing motions,[31] all four positions E11, E15, G8, and G12 appear in the central gate area between cavities DP, Xe4, and Xe2. In Tr. Hb, ligand migration along the protein's internal tunnel is thought to be regulated by residues Gln58-E11 and Phe62-E15, with both side-chains acting as gate-opening molecular switches.[78−81] For human Hb, gating movements of the leucine residue in G12 govern the hopping of gaseous ligand from and to different binding sites.[57]

As we have already seen, most MS residues are also frontier residues adjacent to two or more of the internal cavities that were detected in the various structures produced during our MD simulations. If we look more precisely into the structural rearrangement of our globins inner pockets, it appears clearly that the mechanical variations of the proteins are closely related to the cavity network fluctuations. As an example, we superimposed in Figure 6a and b the five main cavities of Tr. Hb in its THB2 (in blue) and THB4 (in yellow) conformations. The five residues undergoing the most important variations of their force

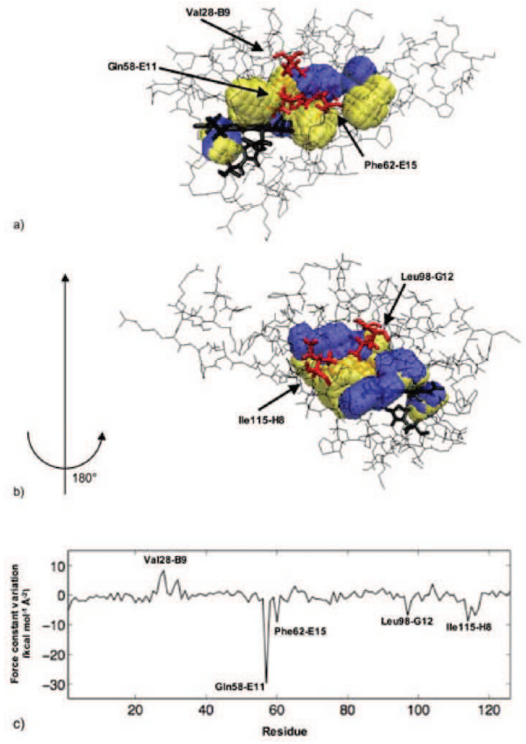


Figure 6. Superposition of the cavity networks from the THB2 (in blue) and THB4 (in yellow) structures from Tr. Hb. The residues undergoing the most important mechanical perturbations upon the structural transition are plotted in red. (a) Front view, (b) back view. (c) Variation of the force constants upon transition from the THB2 to THB4 structures from Tr. Hb.

constant upon the THB2 → THB4 transition are signaled in Figure 6c. Among these are positions E11, E15 (whose importance we underlined in the previous paragraph), and G8 from the mechanical nucleus. Gln58-E11 in particular shows a remarkable decrease of its rigidity (~ -30 kcal mol^{-1} Å^{-2}), which can be related to its central position in the globin's structure. As we can see in Figure 6a and b, the five MS residues from Figure 6c, which are drawn here in red, tightly surround the internal cavities of Tr. Hb. Gln58-E11 and Phe62-E15 lie right at the frontiers between three successive pockets leading to the prosthetic heme group that are found in the THB4 structure (in yellow). Hence, for the ligand to access the heme binding site by diffusing along the protein's cavity network, it is essential for the side-chains of these residues to show some flexibility.

CONCLUDING REMARKS

In our previous works on protein mechanics, we did compare rigidity profiles for various protein oxidation or coordination states and were able to relate residues mechanical properties to their role in the protein's functional activity.[30,50,51] Here, we used classical MD simulations to produce several representative clusterized structures for a single protein state, and the mechanical properties of each structure were then studied via coarse-grain Brownian dynamics. By comparing the rigidity profiles of the clusterized structures, we show that these mechanical properties do have a dynamic quality. While the rigidity profile of a protein remains qualitatively the same along time, with its main peaks associated with a given set of amino acids, it can nonetheless present noticeable variation from one structure to the other for a limited number of residues. In the case of globins, the resulting "mechanically sensitive" residues are connected with the breathing motions of the protein and the fluctuations of its internal cavity network. We also note that these residues positions are well conserved along the protein's sequence. In particular, we could identify what we called a mechanical nucleus, formed by positions E11, E15, G8, and G12. Residues occupying these positions have already been shown separately to play a key role for ligand diffusion in Mgb, Tr. Hb, and human Hb using various experimental and theoretical approaches. Here, we suggest that this quartet might actually be essential for the regulation of ligand migration within the cavity network all throughout the whole protein family. More generally, our findings are of interest for the study of the numerous globular proteins that possess internal cavities and channels, such as redox enzymes.[82−84] From a protein engineering perspective, the study of their mechanical properties should bring us valuable information regarding the key residues that could represent potential mutation targets to modulate or improve their enzymatic activity.

ASSOCIATED CONTENT

Supporting Information. Supplementary Tables S1−6 summarizing the 10 main cavities with their volume and lining residues for each of the 30 globin structures (five for each globin chain) that were produced during this study. Supplementary Table 7 giving the percentage of sequence identity for the six globin chains. Complete refs 42 and 84. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author
sacquin@ibpc.fr

REFERENCES

(1) Hardison, R. C. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5675–5679.
(2) Vinogradov, S. N.; Hoogewijs, D.; Bailly, X.; Arredondo-Peter, R.; Gough, J.; Dewilde, S.; Moens, L.; Vanfleteren, J. R. *BMC Evol. Biol.* **2006**, *6*, 31–67.
(3) Vinogradov, S. N.; Hoogewijs, D.; Bailly, X.; Mizuguchi, K.; Dewilde, S.; Moens, L.; Vanfleteren, J. R. *Gene* **2007**, *398*, 132–142.
(4) Vinogradov, S. N.; Moens, L. *J. Biol. Chem.* **2008**, *283*, 8773–8777.
(5) Kakar, S.; Hoffman, F. G.; Storz, J. F.; Fabian, M.; Hargrove, M. S. *Biophys. Chem.* **2010**, *152*, 1–14.
(6) Wajcman, H.; Kiger, L.; Marden, M. C. *C. R. Biol.* **2009**, *332*, 273–282.
(7) Kendrew, J. C.; Dickerson, R. E.; Strandberg, B. E.; Hart, R. G.; Davies, D. R.; Phillips, D. C.; Shore, V. C. *Nature* **1960**, *185*, 422–427.
(8) Bettati, S.; Viappiani, C.; Mozzarelli, A. *Biochim. Biophys. Acta* **2009**, *1794*, 1317–1324.
(9) Frauenfelder, H. *Chem. Phys.* **2010**, *375*, 612–615.
(10) Nienhaus, K.; Nienhaus, G. U. *IUMB Life* **2007**, *59*, 490–497.
(11) Burmester, T.; Hankeln, T. *J. Exp. Biol.* **2009**, *212*, 1423–1428.
(12) Tilton, R. F.; Kuntz, I. D.; Petsko, G. A. *Biochemistry* **1984**, *23*, 2849–2857.
(13) Brunori, M.; Gibson, Q. H. *EMBO Rep.* **2001**, *2*, 674–679.
(14) Hubbard, S. J.; Gross, K. H.; Argos, P. *Protein Eng.* **1994**, *7*, 613–626.

(15) Carugo, O.; Argos, P. *Proteins* **1998**, *31*, 201–213.
(16) Tomita, A.; Kreutzer, U.; Adachi, S.; Koshihara, S.; Jue, T. *J. Exp. Biol.* **2010**, *213*, 2748–2754.
(17) Case, D. A.; Karplus, M. *J. Mol. Biol.* **1979**, *132*, 343–368.
(18) Milani, M.; Pesce, A.; Ouellet, Y.; Ascenzi, P.; Guertin, M.; Bolognesi, M. *EMBO J.* **2001**, *20*, 3902–3909.
(19) Bourgeois, D.; Vallone, B.; Schotte, F.; Arcovito, A.; Miele, A. E.; Sciara, G.; Wulff, M.; Anfinrud, P.; Brunori, M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 8704–8709.
(20) Schotte, F.; Lim, M. H.; Jackson, T. A.; Smirnov, A. V.; Soman, J.; Olson, J. S.; Phillips, G. N.; Wulff, M.; Anfinrud, P. A. *Science* **2003**, *300*, 1944–1947.
(21) Bossa, C.; Anselmi, M.; Roccatano, D.; Amadei, A.; Vallone, B.; Brunori, M.; Di Nola, A. *Biophys. J.* **2004**, *86*, 3855–3862.
(22) Bossa, C.; Amadei, A.; Daidone, I.; Anselmi, M.; Vallone, B.; Brunori, M.; Di Nola, A. *Biophys. J.* **2005**, *89*, 465–474.
(23) Mouawad, L.; Marechal, J. D.; Perahia, D. *Biochim. Biophys. Acta* **2005**, *1724*, 385–393.
(24) Cohen, J.; Arkhipov, A.; Braun, R.; Schulten, K. *Biophys. J.* **2006**, *91*, 1844–1857.
(25) Ceccarelli, M.; Anedda, R.; Casu, M.; Ruggerone, P. *Proteins* **2008**, *71*, 1231–1236.
(26) Nishihara, Y.; Hayashi, S.; Kato, S. *Chem. Phys. Lett.* **2008**, *464*, 220–225.
(27) Elber, R.; Gibson, Q. H. *J. Phys. Chem. B* **2008**, *112*, 6147–6154.
(28) Elber, R. *Curr. Opin. Struct. Biol.* **2010**, *20*, 162–167.
(29) Cohen, J.; Schulten, K. *Biophys. J.* **2007**, *93*, 3591–3600.
(30) Bocahut, A.; Bernad, S.; Sebban, P.; Sacquin-Mora, S. *J. Phys. Chem. B* **2009**, *113*, 16257–16267.
(31) Scorciapino, M. A.; Robertazzi, A.; Casu, M.; Ruggerone, P.; Ceccarelli, M. *J. Am. Chem. Soc.* **2009**, *131*, 11825–11832.
(32) Maguid, S.; Fernandez-Alberti, S.; Ferrelli, L.; Echave, J. *Biophys. J.* **2005**, *89*, 3–13.
(33) Laberge, M.; Yonetani, T. *IUBMB Life* **2007**, *59*, 528–534.
(34) Pesce, A.; Dewilde, S.; Nardini, M.; Moens, L.; Ascenzi, P.; Hankeln, T.; Burmester, T.; Bolognesi, M. *Structure* **2003**, *11*, 1087–1095.
(35) Evans, S. V.; Brayer, G. D. *J. Mol. Biol.* **1990**, *213*, 885–897.
(36) Makino, M.; Sugimoto, H.; Sawai, H.; Kawada, N.; Yoshizato, K.; Shiro, Y. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 671–677.
(37) Fermi, G.; Perutz, M. F.; Shaanan, B.; Fourme, R. *J. Mol. Biol.* **1984**, *175*, 159–174.
(38) Berendsen, H. J. C.; Vanderspoel, D.; Vandrunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.
(39) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.
(40) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
(41) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
(42) Frisch, M. J.; et al. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2003.
(43) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
(44) Li, H. Y.; Elber, R.; Straub, J. E. *J. Biol. Chem.* **1993**, *268*, 17908–17916.
(45) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952–962.
(46) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
(47) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
(48) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
(49) Hendlich, M.; Rippmann, F.; Barnickel, G. *J. Mol. Graphics Modell.* **1997**, *15*, 359.
(50) Sacquin-Mora, S.; Lavery, R. *Biophys. J.* **2006**, *90*, 2706–2717.
(51) Sacquin-Mora, S.; Laforet, E.; Lavery, R. *Proteins* **2007**, *67*, 350–359.
(52) Lavery, R.; Sacquin-Mora, S. *J. Biosci.* **2007**, *32*, 891–898.
(53) Zacharias, M. *Protein Sci.* **2003**, *12*, 1271–82.
(54) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–50.
(55) Ermak, D. L.; McCammon, J. A. *J. Chem. Phys.* **1978**, *69*, 1352–1360.
(56) Pastor, R. W.; Venable, R.; Karplus, M. *J. Chem. Phys.* **1988**, *89*, 1112–1127.
(57) Savino, C.; Miele, A. E.; Draghi, F.; Johnson, K. A.; Sciara, G.; Brunori, M.; Vallone, B. *Biopolymers* **2009**, *91*, 1097–1107.
(58) Vallone, B.; Nienhaus, K.; Brunori, M.; Nienhaus, G. U. *Proteins* **2004**, *56*, 85–92.
(59) Ptitsyn, O. B.; Ting, K. L. *J. Mol. Biol.* **1999**, *291*, 671–82.
(60) Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. *Bioinformatics* **2007**, *23*, 2947–2948.
(61) Maguid, S.; Fernandez-Alberti, S.; Parisi, G.; Echave, J. *J. Mol. Evol.* **2006**, *63*, 448–457.
(62) Hollup, S. M.; Fuglebakk, E.; Taylor, W. R.; Reuter, N. *Protein Sci.* **2010**, *20*, 197–209.
(63) Olson, J. S.; Soman, J.; Phillips, G. N. *IUBMB Life* **2007**, *59*, 552–562.
(64) Vallone, B.; Nienhaus, K.; Matthes, A.; Brunori, M.; Nienhaus, G. U. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17351–17356.
(65) Anselmi, M.; Brunori, M.; Vallone, B.; Di Nola, A. *Biophys. J.* **2007**, *93*, 434–441.
(66) Lutz, S.; Nienhaus, K.; Nienhaus, G. U.; Meuwly, M. *J. Phys. Chem. B* **2009**, *113*, 15334–15343.
(67) Abbruzzetti, S.; Faggiano, S.; Bruno, S.; Spyrakis, F.; Mozzarelli, A.; Dewilde, S.; Moens, L.; Viappiani, C. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 18984–18989.
(68) Nienhaus, K.; Lutz, S.; Meuwly, M.; Nienhaus, G. U. *ChemPhysChem* **2010**, *11*, 119–129.
(69) Anselmi, M.; Di Nola, A.; Amadei, A. *J. Phys. Chem. B* **2011**, *115*, 2436–2446.
(70) de Sanctis, D.; Dewilde, S.; Pesce, A.; Moens, L.; Ascenzi, P.; Hankeln, T.; Burmester, T.; Bolognesi, M. *J. Mol. Biol.* **2004**, *336*, 917–927.
(71) de Sanctis, D.; Dewilde, S.; Pesce, A.; Moens, L.; Ascenzi, P.; Hankeln, T.; Burmester, T.; Bolognesi, M. *Biochem. Biophys. Res. Commun.* **2004**, *316*, 1217–1221.
(72) Orlowski, S.; Nowak, W. *BioSystems* **2008**, *94*, 263–266.
(73) Crespo, A.; Marti, M. A.; Kalko, S. G.; Morreale, A.; Orozco, M.; Gelpi, J. L.; Luque, F. J.; Estrin, D. A. *J. Am. Chem. Soc.* **2005**, *127*, 4433–4444.
(74) Golden, S. D.; Olsen, K. W. *Globins and Other Nitric Oxide-Reactive Proteins, Part B*; Elsevier Academic Press, San Diego, CA, 2008; Vol. 437, pp 417−437.
(75) Quillin, M. L.; Li, T. S.; Olson, J. S.; Phillips, G. N.; Dou, Y.; Ikedasaito, M.; Regan, R.; Carlson, M.; Gibson, Q. H.; Li, H. Y.; Elber, R. *J. Mol. Biol.* **1995**, *245*, 416–436.
(76) Ishikawa, H.; Uchida, T.; Takahashi, S.; Ishimori, K.; Morishima, I. *Biophys. J.* **2001**, *80*, 1507–1517.
(77) Dantsker, D.; Roche, C.; Samuni, U.; Blouin, G.; Olson, J. S.; Friedman, J. M. *J. Biol. Chem.* **2005**, *280*, 38740–38755.
(78) Bidon-Chanal, A.; Marti, M. A.; Crespo, A.; Milani, M.; Orozco, M.; Bolognesi, M.; Luque, F. J.; Estrin, D. A. *Proteins* **2006**, *64*, 457–464.
(79) Bidon-Chanal, A.; Marti, M. A.; Estrin, D. A.; Luque, F. J. *J. Am. Chem. Soc.* **2007**, *129*, 6782–6788.
(80) Mishra, S.; Meuwly, M. *Biophys. J.* **2009**, *96*, 2105–2118.
(81) Lama, A.; Pawaria, S.; Bidon-Chanal, A.; Anand, A.; Gelpi, J. L.; Arya, S.; Marti, M.; Estrin, D. A.; Luque, F. J.; Dikshit, K. L. *J. Biol. Chem.* **2009**, *284*, 14457–14468.
(82) Fontecilla-Camps, J. C.; Amara, P.; Cavazza, C.; Nicolet, Y.; Volbeda, A. *Nature* **2009**, *460*, 814–822.

(83) Baron, R.; Riley, C.; Chenprakhon, P.; Thotsaporn, K.; Winter, R. T.; Alfieri, A.; Forneris, F.; van Berkel, W. J. H.; Chaiyen, P.; Fraaije, M. W.; Mattevi, A.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 10603–10608.
(84) Liebgott, P. P.; et al. *Nat. Chem. Biol.* **2010**, *6*, 63–70.
(85) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 27–8.